



Education
Endowment
Foundation

Lesson planning using AI lesson assistant, Aila - A Teacher Choices Trial

Evaluation report

October 2026

Palak Roy, Helen Poet, Ruth Staunton, Katherine Aston,
David Thomas and Ben Styles



NFER

National Foundation for
Educational Research



The Education Endowment Foundation is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, nurseries and colleges to improve teaching and learning for 2 – 19-year-olds through better use of evidence.

We do this by:

- **Summarising evidence.** Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds.
- **Putting evidence to use.** Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were re-endowed with an additional £137m, allowing us to continue our work until at least 2032.

For more information about the EEF or this report please contact:



Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP



0207 802 1653



info@eefoundation.org.uk



www.educationendowmentfoundation.org.uk

Contents

About the evaluator	5
Executive summary	6
Introduction	8
Methods	19
Impact evaluation results.....	43
Implementation and process evaluation results	64
Conclusion.....	91
References.....	97
Appendix A: Security classification of trial findings	99
Appendix B: Tables and Charts for impact evaluation results.....	101
Further appendices	105

About the evaluator

The project was independently evaluated by a team from the National Foundation for Educational Research (NFER). The trial director and principal investigator for this study was Helen Poet (Research Director). Palak Roy (Senior Trials Manager) led the evaluation team including the impact evaluation. The implementation and process evaluation was led by Katherine Aston (Research Manager) supported by Daniel Jackson (Educational Assessment Developer) as the implementation and process evaluation researcher. David Thomas (Senior Research Manager) led the lesson and resource quality strand. Ruth Staunton (Senior Statistician) was the trial statistician, leading the quantitative analysis, with additional analytical support provided by Jose Liht (Senior Statistician). Julie Thompson supported the team with administrative tasks. They were supported by NFER's research operations team for the collection of evaluation data and liaising with participants. The NFER research operations team included Kathryn Hurd, Jishi Jose, Rob Green, Andrei Illes, and Elaine McCann.

Contact details

Palak Roy
The Mere, Upton Park
Slough
Berkshire. SL1 2DQ

t: + 44 (0) 1753 574123

e: p.roy@nfer.ac.uk

Acknowledgements

We are grateful to all the primary schools and teachers who took part in this evaluation. We appreciate the time and insight offered by teachers who contributed through interviews and in-person case study visits.

We would also like to acknowledge colleagues at the EEF, including Faizaan Sami and Fabiola Clemente, for their continued guidance and support. Our thanks also extend to members of the Study Advisory Board—Libby Hills, Chris Goodall, Ria Drabble, Michael Adkins, and Jake Anders—for their valuable input.

We are also grateful to colleagues at Oak National Academy, including Hannah-Beth Clark, Reka Budai, Laura Benton, Will Gayne, Ewan Wakeman, Emma Searle, and Rachel Urquhart for their contributions during the set-up phase.

We would also like to extend our thanks to colleagues at NFER—Daniel Jackson, Kathryn Hurd, Jishi Jose, Rob Green, Andrei Illes, Elaine McCann, Jose Liht and Julie Thompson—whose support was instrumental throughout the project.

Executive summary

The project

Oak National Academy’s Aila is an AI-powered lesson assistant designed to support teachers with lesson planning and resource creation. Aligned to the national curriculum for England, Aila helps teachers generate lesson materials by specifying a lesson objective and refining the lesson content to meet their needs.

The trial aimed to evaluate the impact of Aila on teachers’ lesson planning workload compared to a non-Aila group. Participating teachers in Years 3, 4, 5, and 6 implemented their allocated approaches over ten weeks during autumn term 2025. The two-arm cluster randomised controlled trial randomly allocated 464 teachers from 108 schools, at school level, to either Aila or non-Aila arms. Teachers in the Aila group were encouraged to use Aila while those in the non-Aila group were asked not to use it. Both groups could use other Generative Artificial Intelligence (GenAI) resources (such as ChatGPT or Microsoft Copilot) but were asked not to use the Oak National Academy website to find new resources to isolate the effect of using Aila. Teachers in both groups received guidance about the condition they were randomised to, including ‘dos and don’ts’.

The trial was an Education Endowment Foundation (EEF) funded Teacher Choices trial. Teacher Choices trials aim to test everyday choices educators make when planning lessons and supporting pupils. The primary outcome measure was the average time spent planning Key Stage 2 (KS2) lessons over ten weeks, measured using a weekly teacher diary. The secondary outcome measure was an assessment of lesson resource quality by six independent teacher-reviewers. The implementation and process evaluation (IPE) examined usual lesson planning practices, adherence to assigned approaches, how and why teachers used Aila, and perceived impacts on workload and teaching. The trial started in June 2025 and concluded in February 2026.

Table 1: Key conclusions

Key conclusions	
1.	The lesson and resource preparation time spent by the Aila group was 76% of the time spent by the non-Aila group during the ten weeks of the trial. In this trial, teachers in the non-Aila group spent 3.3 hours per week preparing lessons and resources for their class, while teachers in the Aila group spent around 2.5 hours per week, saving 0.8 hours (or 48.6 minutes) per week on average.
2.	This reduction in teacher workload was evident without compromising the lesson resource quality: the lesson resource quality analysis showed that the quality of the lesson resources in the Aila group was no worse than that in the non-Aila group.
3.	Teachers in the Aila group with lower confidence in their subject knowledge, lower pedagogical knowledge, those new to the profession or with higher confidence in technology were more likely to have a lower workload than similar teachers in the non-Aila group.
4.	Teachers in the Aila group saved time even though Aila was used for an average of 7% of lessons. Notably, usage frequency and amount both declined over the ten weeks.
5.	Teachers had mixed views about the Aila tool: while they found it to be accessible and curriculum-aligned, many preferred alternative GenAI tools which provided more ease of use, and Aila, by design, seeks to assist teachers through the planning process, requiring more involvement. This may also reflect a disconnect between Aila’s intended purpose of supporting lesson planning from scratch and teachers’ planning practices as well as perceptions that its outputs often required adaptation before being classroom ready.

EEF security rating

These findings have a high security rating. This was a Teacher Choices trial, testing an everyday choice teachers make when planning and preparing lessons. The trial was a well-designed, two-armed randomised controlled trial. The trial was well-powered. Due to some teachers not completing the weekly diary data, 10.3% teachers were not included in the primary analysis.

Additional findings

The time teachers spent on lesson planning in the Aila group was 76% of the time spent on planning in the non-Aila group. This is our best estimate of impact. As with any study, there is some uncertainty around the result: the reduction in workload could plausibly lie between 64% and 89%. However, the evidence indicates a clear reduction in teacher workload. This is similar to the findings from the ChatGPT trial where teachers in the ChatGPT group spent 69% as much time on lesson planning as the non-GenAI group (Roy et al., 2024). The time savings did not compromise the quality of lesson resources delivered during the trial.

Teachers who used Aila more frequently experienced greater reduction in lesson planning time, a pattern that remained when excluding the familiarisation week and half-term break. Before the trial, intended planning levels were similar across both groups. Most teachers adhered to their assigned approach, although adherence was somewhat lower in the Aila group. Overall, Aila use was relatively low: 81% of Aila teachers used it at least once during the trial, but it was used for only 7% of lessons in an average week and declined over time. It was used mainly for planning whole lessons rather than producing stand-alone resources. Teachers typically drew on a range of planning approaches, including existing resources and other GenAI tools, meaning Aila was often used as a supplementary tool rather than replacing existing practices. Most teachers also adapted Aila's outputs, either within the tool or after downloading resources, as intended. Despite this relatively low and selective use, using Aila led to a meaningful reduction in teacher workload.

Some teachers reported benefits: they valued being able to use Aila flexibly and found it useful for planning lessons on unfamiliar topics. Others found Aila outputs generic or inflexible, requiring additional adaptation, and fewer than one in five felt Aila meaningfully influenced their teaching and planning. Many preferred other tools or existing approaches. This may reflect a disconnect between Aila's intended purpose of supporting lesson planning from scratch and teachers' day-to-day planning needs. In line with this, actual use of Aila in the spring term following the trial was lower than originally anticipated.

These findings broadly support parts of the logic model, particularly the assumption that Aila can reduce planning time while producing acceptable-quality resources. However, uptake was limited because teachers did not often need to plan whole lessons from scratch and many felt further adaptation was required. As a result, Aila was more commonly used alongside existing approaches than as a replacement. Overall, the findings suggest that Aila did not consistently offer advantages over teachers' existing approaches and did not always align with their planning preferences and needs.

These findings are important given that almost all teachers in this trial reported lesson planning as a source of stress and believed they spend too much time on it, consistent with wider evidence on teacher workload and time pressures (IFF Research and UCL Institute of Education, 2025; McLean and Worth, 2025). They also highlight the potential for AI tools to alleviate workload while underlining the importance of usability, flexibility, and alignment with teacher needs, and their autonomy to draw on whichever sources best fit their individual circumstances.

Impact

Table 2: Summary of impact on primary outcome(s)

Outcome/ Group	Lesson planning time for Aila group expressed as a % of the time for non-Aila group (95% confidence interval)	EEF security rating	Adjusted mean for Aila group (95% CI)	Adjusted mean for non-Aila group (95% CI)	No of teachers (missing)	p-value
RQ1: Weekly hours spent in lesson and resource preparation	76% (64%, 89%)		2.51 (2.21, 2.86)	3.32 (2.9, 3.79)	416 (48)	0.001

Introduction

Overview of evaluation design

The evaluation was commissioned in response to the rapidly increasing use of GenAI by teachers: 42% of primary and secondary teachers reported using GenAI in their role (DfE and OIT, 2024), which increased to 58% by the time the study started (Teacher Tapp, 2025c). This reflects a clear upward trend in both familiarity with the tools and their usage. Teachers most commonly report using GenAI tools to support lesson planning, resource creation, and administrative activities, with planning-related uses most prevalent: over half of teachers reporting use for this purpose (Teacher Tapp, 2025b). Emerging evidence suggests that this increased GenAI adoption in lesson planning can make a valuable contribution to teachers' planning workload, reducing time spent on this activity by 31% (DSIT et al., 2024; Roy et al., 2024).

Despite this growth, the use of education-specific AI tools remains limited. While around a third of teachers surveyed by Teacher Tapp in autumn 2024 reported using a generic AI tool (such as ChatGPT) 'within the last week' (34%), the same research found that only 3% had used an education-specific AI-product in the previous week. Teachers reported using these tools to generate resources such as lesson plans, worksheets, model answers, comprehension questions, and assessments. Many also said they used GenAI tools to differentiate learning materials for students (Teacher Tapp, 2025a). While 'generic' GenAI can have many uses, users need to ensure that the content is aligned and appropriate for the curriculum and context that they teach in—this can include tailoring to the English school context and National Curriculum, but also what may seem like minor points such as region-specific spelling (U.K. rather than American, for example).

In response to these concerns, the U.K. government announced an initiative to create a high-quality education content store to make AI outputs more reliable for educators in England and to support the training of GenAI tools tailored for England's school system and National Curriculum (DSIT et al., 2024). In contrast to general purpose GenAI like ChatGPT, Microsoft Copilot, or Claude, which lack curriculum alignment with England's primary school which uses National Curriculum, this initiative focuses on building education-specific GenAI that is trained on national guidelines, lesson plans, and assessments. Alongside this, the Department for Education (DfE) released its first guidance promoting the safe and effective use of GenAI to reduce workload, particularly in planning, marking, and administrative tasks (DfE, 2025a). While initiatives such as the DfE's EdTech Evidence Board (DfE, 2025a) and the Content Store (DSIT et al., 2024) represent positive steps, they are still in development or at pilot stage and it is too early to determine their potential long-term role or impact. In this context, building a strong evidence base on the outcomes of teachers' or pupils' use of AI in education remains important. Nevertheless, these initiatives demonstrate an important move towards the creation of curriculum-aligned AI tools tailored to the English education system (Shearing, 2025).

Aila, developed by Oak National Academy, represents one such prominent example of education-specific GenAI tools. In 2024 Oak launched the beta version of Aila, which walks teachers through the process of creating a suite of resources for a lesson: to create a lesson, teachers need to input the title, key stage, and subject and to create teaching material they need to specify the year group, subject, and lesson title or learning outcome. They then interact with Aila to refine the focus and content of the lesson according to their needs. There are also opportunities for the teacher to specify the age range and any needs of the pupils (including any special educational needs, different reading ages, prior knowledge/content covered with the class). At the end of the process, Aila creates a lesson plan, slide deck, quizzes, and worksheets that can be downloaded and edited further by the teacher (for example, Aila does not currently include images so teachers can supplement the downloaded slide deck with relevant pictures). As such, Aila appears to be most useful for teachers who know which lesson they need to plan but who do not have any resources already created.

Aila uses education resources developed by a range of educators and subject experts, which are used to inform the lessons and resources it generates (Strom, 2025). Users interact with Aila through a chat interface powered by a large language model (LLM), which at the time of the trial was OpenAI's GPT-4o. The tool has been customised via

prompt engineering and retrieval-augmented generation (RAG) combined with content anchoring, which allows the model to prioritise accessing material from Oak National Academy's corpus of education content in the first instance. Where the requested content is not available within that corpus, the LLM generates new content within safety guardrails that Oak National Academy says help ensure it is 'safe and high-quality'. These guardrails are intended to help Aila produce age-appropriate, English National Curriculum-aligned lesson content by guiding its responses, checking inputs and outputs against clearly defined safety categories. Teachers review and refine materials before classroom use (Clark, 2025). Aila is free for U.K. teachers, requiring only an account to log in.

Within just two months of its release, Aila was used by more than 10,000 educators from all key stages to generate nearly 25,000 lesson plans, predominantly in maths and English (Benton, 2025). User feedback collected by Oak National Academy was positive, with 85% (out of 3,301 survey respondents) rating the structure and content of the lesson plans as high or very high quality. Nearly two-thirds of teachers reported that Aila reduced their workload or freed time for other tasks, saving between one and 15 hours. Many noted reduced stress and improved wellbeing, especially when creating lesson plans from scratch. However, the study also identified areas for further development, including expanding the curriculum content underpinning Aila, improving editing functionality, and adding visual elements such as images to generated slide decks. These findings led the EEF to commission an evaluation of Aila's use for lesson planning through the Teacher Choices programme.

About this project

This was a Teacher Choices trial—one testing an everyday choice teachers make when planning their lessons. They aim to generate practical, actionable evidence on common teaching practices, helping teachers, school leaders, and policymakers make informed decisions and to improve the relevance and feasibility of education research in real classroom settings.

This evaluation employed a two-arm cluster randomised controlled trial, with randomisation at the school level. Schools were allocated to one of two approaches: Aila and non-Aila. Teachers in the Aila group were encouraged to use Aila to support lesson planning and those in the non-Aila group were asked not to use Aila (Aila is freely available to U.K. teachers). Both groups were asked not to use the Oak National Academy website for resources during the same period due to planned developments by Oak National Academy to further integrate AI tools in the main website during the trial period. Teachers in both groups were allowed to use other GenAI tools (for example, ChatGPT, Dall-E, Gemini). Restricting their use would not reflect real-world practice or genuine Teacher Choices, especially given their use is becoming increasingly common and the Department for Education was recommending teachers integrate it into their practice. As such, there was no restriction on the use of other GenAI tools in either group, including other education-specific AI tools. (See the next section for further details on the trial arms and choices.)

Participating class teachers in Years 3, 4, 5, and 6 (KS2) implemented their allocated approach over a ten-week period over the autumn term in 2025. Randomisation at school level was chosen to minimise the risk of contamination between teachers within the same school, particularly given the prevalence of collaborative planning practices within schools.

Outcomes were measured at the teacher level. The primary research question was: what is the impact on workload of using Aila for lesson planning, compared to a non-Aila group who will be encouraged not to use Aila? Since both groups were allowed to use other GenAI tools, the non-Aila group was considered aligned to business as usual (with the exception of any use of Aila). Thus, the primary outcome was the average time spent on lesson planning per week, during the ten-week trial period, including any planning that took place during the autumn half-term break. The baseline measure was the expected weekly time spent on lesson planning, collected in the baseline survey before randomisation.

The secondary research question examined any differences in the quality of lesson plans and resources produced by the Aila group and the non-Aila group. The secondary outcome was the quality of lesson and resource materials prepared during the ten-week trial period measured via the average score given by an independent panel of teachers. There were several additional research questions, which are listed in the Evaluation Objectives section.

The implementation and process evaluation (IPE) was conducted alongside the impact evaluation to explore how the approaches were implemented in practice and to provide contextual information to support interpretation of the impact findings. Drawing on multiple data sources, including teacher surveys, weekly diaries, interviews, case study visits, observations, and Aila usage data, the IPE examined usual lesson planning practice, adherence to assigned approaches, how and why teachers used Aila, and perceived impacts on workload and teaching. It also characterised teachers' capability, opportunity, and motivation (see 'behaviour change framework'; Michie, van Stralen and West, 2011) and explored perceived impacts for both teachers and pupils. The IPE research questions are summarised in the Evaluation Objectives section (with sub-questions presented in the IPE Methods section) and were designed to support interpretation of impact findings in relation to teacher practice.

Scoping activities

Aila is designed to support teachers who are creating lessons from scratch, although it assumes that teachers will typically be working within an existing scheme of work or curriculum framework. To assess the feasibility of the trial and refine eligibility criteria, the light-touch scoping phase focused on understanding three key aspects of current practice in KS2: (i) the extent to which teachers plan lessons from scratch versus using pre-prepared materials, (ii) whether planning is conducted individually or collaboratively, and (iii) the prevalence of GenAI use specifically for lesson planning. Establishing these baseline practices was necessary to determine both the potential demand for the tool and the size of the eligible participant pool, as well as to characterise 'business-as-usual' conditions.

To address these questions, a set of three survey items was commissioned through Teacher Tapp in March 2025. The survey received responses from more than 2,500 teachers in state-maintained primary schools in England, providing a robust snapshot of current practice. These findings complement wider evidence from Teacher Tapp (Teacher Tapp, 2025a), which indicated that 34% of teachers had used a generic AI tool—such as ChatGPT—in the previous week, although this did not distinguish between different use cases such as lesson planning or administrative tasks.

Findings from the first survey question, which asked teachers how most of their lessons were planned, indicated that lesson planning in KS2 involves a combination of individual and collaborative approaches. Over a third (35%) reported independently planning most of their lessons, while a further 29% reported planning collaboratively to create shared resources. Collaborative planning was more prevalent in larger schools. These findings suggested that a substantial proportion of teachers were directly engaged in lesson planning, either independently or jointly, and therefore had opportunities to use tools such as Aila. This finding also informed the choice of school-level randomisation in the trial design, given the potential for contamination where collaborative planning is common.

The survey also explored the degree of autonomy teachers had in selecting resources. More than half of respondents (54%) indicated that they had total or mostly free choice over the resources they used in teaching, while only 3% reported relying entirely on pre-prepared resources across all subjects. This indicates that most teachers had discretion in their planning approaches, suggesting that there was sufficient flexibility within existing practice for teachers to adopt a new tool such as Aila, supporting the feasibility of recruitment by indicating that a sizeable proportion of teachers would be able to choose whether or not to use Aila.

Given the rapidly evolving use of GenAI in education, the scoping phase also examined the prevalence of AI use specifically for lesson planning. When asked how often they used AI tools—such as ChatGPT or Gemini—when preparing lessons or resources, 45% of teachers reported doing so at least a few times per month, including 17% who reported weekly use. These findings suggested that, while GenAI was already used by a substantial minority of teachers, its use for lesson planning was not yet universal. This finding aligned with broader survey evidence indicating increasing uptake of GenAI tools (Teacher Tapp, 2025a) while also highlighting variation in frequency of use across teachers.

Taken together, the scoping findings indicated that GenAI use for lesson planning had not yet reached saturation, suggesting that the trial was likely to be viable in terms of recruitment and contrast between trial arms. At the same time, the observed variation in prior use informed the inclusion of a sensitivity analysis based on teachers' existing use of GenAI in the trial design.

Full results from the scoping phase are attached in Appendix C.

Teacher Choice approaches

The trial focused on Aila as an education-specific GenAI tool to support lesson planning. Aila offers features tailored to classroom use that are expected to make it more effective than general-purpose tools like ChatGPT (Oak National Academy, 2025) by doing the following:

- it intentionally pauses at key points in the process, to encourage the use of teacher expertise in guiding, refining, and checking Aila's outputs;
- it uses retrieval augmented generation to prioritise drawing on existing Oak National Academy curriculum plans and resources, which have been constructed and quality-assured by subject experts and are aligned to the (DfE, 2014); and
- it incorporates pedagogical design into its prompts, including Oak National Academy's curriculum principles, lesson planning guidance, and examples and non-examples.

Aila offers two functions: creating a lesson and creating teaching materials. The function to create teaching materials was added to Aila in August 2025, after study design and participant recruitment but before randomisation and the trial period. Minor changes to the IPE research questions were made to reflect this change.

Using Aila to create a lesson plan begins with the teacher selecting a year group, subject, and lesson topic/title. Guided by the teacher, Aila produces a learning outcome, divides this into smaller learning cycles, and then suggests explanations and activities to use in the lesson. There are regular prompts for the teacher to check and (if needed) adjust the content and focus of the lesson. Aila generates a suite of materials for the lesson, including a slide deck, quizzes, worksheets, and a lesson plan. As a default, generated resources are pitched for the ability of an average pupil in the relevant year group, and teachers can request a higher or lower reading age for the resource. They can also tailor the lesson content by requesting adaptations. For example, a teacher in Manchester could prompt Aila to adapt a lesson on the industrial revolution to their local area. This function is particularly suited to lesson planning 'from scratch', that is, planning based on a lesson title, learning objective, or learning outcome rather than being a tool to supplement lessons already written and in use (though teachers could use it in this way if they wished).

The more recent function in Aila (added in August 2025) to create stand-alone teaching materials allows the user to create either a starter quiz, an exit quiz, a glossary, or a comprehension task. The teacher selects a year group, subject, and lesson title or outcome. Aila generates lesson details and checks these with the teacher before generating the requested resource. This function is particularly suited to supplementing a pre-developed plan with additional teaching materials.

While one group were encouraged to use Aila in their lesson preparations, we carefully considered what the other group should be and decided that the most useful comparison for teachers was to compare the use of Aila with no use of Aila. We expected that this latter group would include a variety of lesson planning approaches, including teachers who use other AI tools (education-specific or generic) as well as teachers who do not use any AI. This would offer a useful 'real world' contrast given the increasing prevalence of GenAI in lesson planning.

As Aila prioritises drawing on Oak National Academy curriculum plans and resources (including 'content anchoring', which means incorporating content from the closest Oak National Academy lesson), there are strong similarities between Aila-generated resources and other Oak National Academy resources. Oak National Academy is constantly developing and improving its offer for teachers and, indeed, the role of Aila within its website evolved during the course of the trial period, for example, with the addition of the option to create stand-alone teaching materials for each lesson on its main website using AI (specifically Aila). To isolate the effect of using Aila (rather than using Oak National Academy resources), to guard against potential contamination and to minimise the impact of any changes that may occur on the platform during the trial period, we asked teachers in both groups to avoid

downloading Oak National Academy resources (monitored via self-report only). Teachers were allowed to continue using any Academy resources they used in previous terms, in the same way that they were permitted to use other pre-existing lesson plans/resources (from previous academic term) that they would have re-used in the absence of the trial.

We monitored any use of Aila in both groups through the weekly diaries (see the Methods section for further details). We conducted subgroup analysis to explore the effect of Aila for teachers with no or low previous experience of using GenAI in lesson planning.

Brief trial guidance, including ‘do’s and don’ts’ for the two arms, were produced by the research team and shared with teachers along with their school’s random allocation (see Appendix D for the guidance documents). Teachers in the Aila arm were asked to register on the Aila platform and use Aila for their KS2 lesson planning for ten weeks in the autumn term 2025. The guidance also included links to a walk-through video and a brief how-to guide produced by Oak National Academy. Teachers in the non-Aila arm were asked specifically not to use Aila when planning KS2 lessons during the same period. The trial guidance documents reminded teachers not to access new Oak National Academy resources in either group during the trial, while allowing use of previously downloaded Oak National Academy materials and other existing planning resources. The TiDieR framework (Hoffmann et al., 2014) for this trial is outlined in Table 3.

Table 3: Intervention Description

	Aila	Non-Aila
Why (rationale)	Research has shown that GenAI tools can save teachers time in lesson and resource preparation. As an education-specific GenAI tool, Aila was expected to show enhanced effects compared with generic tools, including reducing teacher time in lesson planning, developing teachers’ planning capabilities, and maintaining or improving the quality of planned lessons.	The non-Aila group were encouraged to refrain from using Aila, to avoid contamination. We expected activity to reflect the wide variety of lesson planning approaches found in KS2, and so to provide a ‘real world’ comparison of the effect of using Aila.
Who (recipients)	Teachers of Key Stage 2 classes (Years 3 to 6) in English schools.	
What (materials)	Access to Aila, which includes in-built support and prompts for teachers; brief trial guidance (Appendix D), including signposting to Aila sign-up; and guidance on the Oak National Academy website (all free to access).	Brief trial guidance (Appendix D).
What (procedures)	Encouraged to use Aila to support their lesson planning, as much as they could; asked not to access or download any Oak National Academy lessons during the autumn term, except via Aila; and all teachers were asked to plan the number of lessons they would usually plan, rather than planning for additional lessons (for example, planning further ahead).	Asked to continue with their previous lesson planning approach; asked not to use Aila to support their lesson planning; and asked not to access or download any Oak National Academy lessons during the autumn term. All teachers were asked to plan the number of lessons they would usually plan, rather than planning for additional lessons (for example, planning further ahead).
How (delivery Guide)	After randomisation, teachers received the trial guidance for their randomly allocated approach (Appendix D) so that they could familiarise themselves with that approach.	
Where (location)	State-funded mainstream primary schools in England (Years 3 to 6).	

When and how much (dosage)	The trial took place in ten weeks in the autumn term (September to December) 2025. Teachers were encouraged to use Aila to support their lesson planning for KS2 lessons as much as they want over a ten-week period.	The trial took place in ten weeks in the autumn term (September to December) 2025. Teachers were asked not to access the Oak National Academy website—including Aila—for any KS2 planning over a ten-week period.
Tailoring (adaptation)	Teachers were encouraged to use Aila to support planning according to need and preference, for example, for particular subjects, lessons, and/or year groups.	Teachers were asked to continue using their usual planning approach. They were free to use any approach, (including use of GenAI) except for using Aila or the Oak National Academy website.

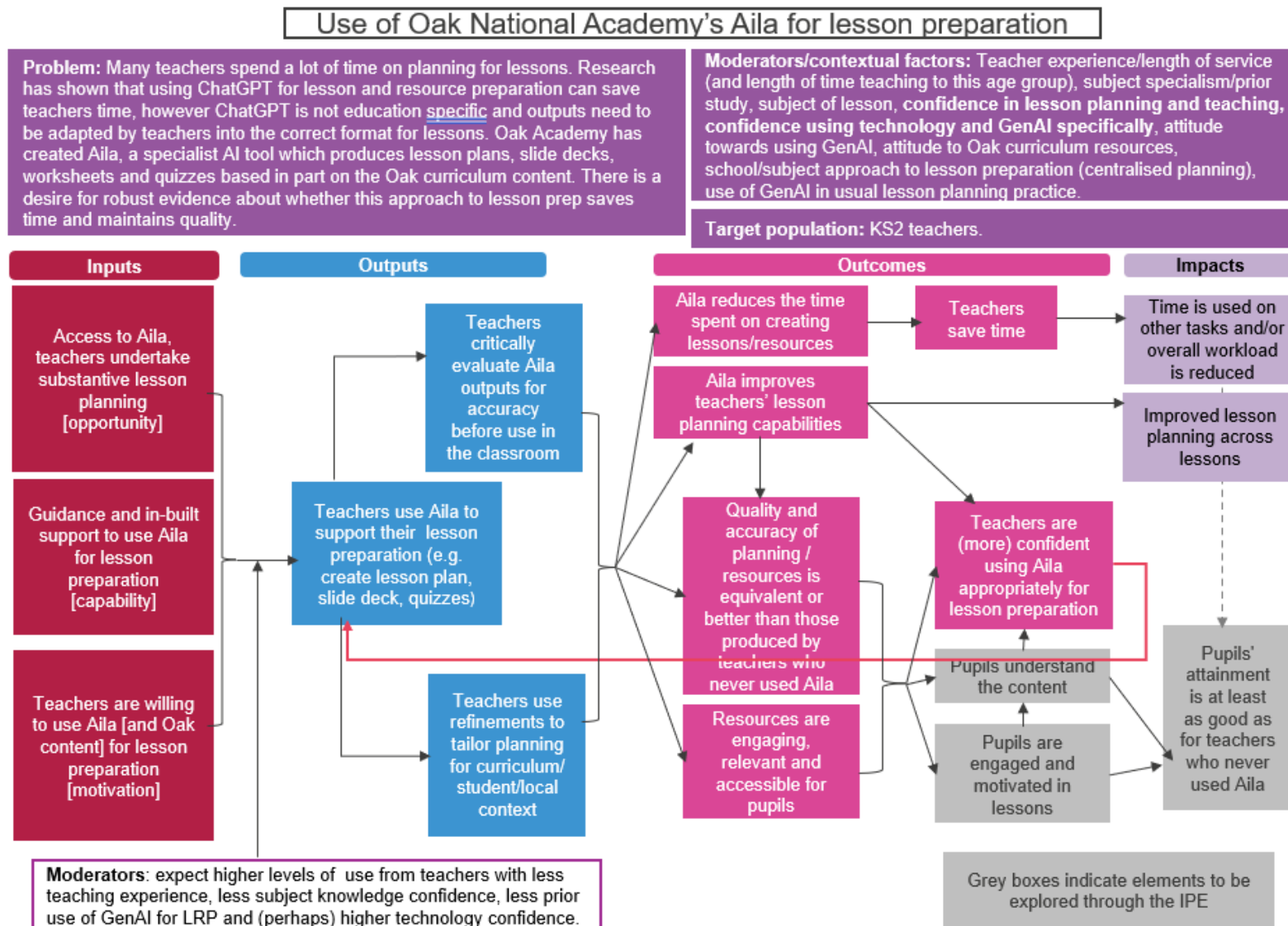
We constructed a logic model for Aila during the project start-up phase, shown in Figure 1. We used the COM-B approach (Michie, van Stralen and West, 2011) to identify the inputs (COM) that lead to the behaviour change (B: the use of Aila). Oak National Academy has identified three intended outcomes from using Aila compared with other lesson planning approaches: improved lesson planning capabilities, improved lesson quality, and reduced time spent on lesson planning.

The core components that Oak National Academy view as distinguishing Aila from other lesson planning approaches are:

- a scaffolded process of lesson planning that aims to model expert lesson planning, which we expected to improve lesson quality and develop teachers' lesson planning capabilities;
- a conceptual framework for pedagogy drawing on cognitive load theory (Sweller, 1988, 2005), Mayer's 'theory of multimedia learning' (Mayer, 2005), and Rosenshine's 'principles of instruction' (Rosenhine, 2012)—a framework included in Aila's underlying prompts to guide the AI generation of lesson plans and resources;
- quick and automatic production of resources in response to teacher inputs (compared with non-GenAI approaches);
- quick and automatic adaptation and personalisation of resources (compared with non-GenAI approaches), which we expected to save planning time and/or improve the tailoring of a lesson for the needs of a specific class, for example, pupil attainment or learning needs; and
- retrieval augmented generation, which prioritises drawing on existing Oak National Academy resources: as these resources were developed and quality-assured by subject experts, we expected lesson plan/resource quality to be maintained or improved compared with other planning approaches, including compared with lesson plans and resources generated by broader GenAI tools that are not inherently tailored to the national curriculum in England or to U.K. English.

It is also worth noting that participants who were subscribed to Oak National Academy's general newsletter may have received additional marketing emails highlight time-saving benefits—you could 'save five hours a week' and 'it could save you hours'. While this could have influenced teachers' perceptions and self-reported measures of saving planning time (the primary outcome for this trial), we have no direct evidence that this occurred. Newsletter distribution varied by topic and relevance, meaning not all subscribers received all emails and only a subset of Aila users who would have subscribed to Oak National Academy's generic communications would have received this.

Figure 1: Logic model



Evaluation objectives

The aim of this evaluation was to assess the impact of using Aila, an AI-powered lesson planning assistant, on teachers' workload and the quality of lesson planning in Key Stage 2 as well as to understand how the tool was used in practice and its perceived effects. A full study plan (Roy et al., 2024) is available on the EEF website.¹

Impact evaluation research questions

Primary research question

RQ1 (Primary RQ): What is the impact of Aila on the average time spent per week on lesson planning for KS2 lessons during the ten-week trial period?

Secondary research question

RQ2: How does the quality of lesson plans and resources in the Aila group compare to the quality of lesson plans and resources in the non-Aila group?

Sensitivity analyses

RQ3: What is the impact of Aila on average time spent per week on lesson planning for KS2 lessons, excluding the first week of Aila use (to allow for familiarisation)?

RQ4: What is the impact of Aila on average time spent per week on lesson planning for KS2 lessons, excluding half-term week(s)?

Subgroup analyses

RQ5a: What is the impact of Aila on the average time spent per week on lesson planning among teachers who indicate low or no prior use of GenAI in lesson planning?

RQ5b: What is the impact of Aila on the average time spent per week on lesson planning among teachers who indicate low or no prior use of GenAI and who intend to plan a high proportion of lessons from scratch?

Complier Average Causal Effect (CACE)

RQ6: What is the impact of using Aila on the average time spent per week on lesson planning among teachers who comply with their allocated approach?

Implementation and process evaluation research questions

IPE RQ1: What is usual practice in lesson planning in KS2 for teachers in the trial?

IPE RQ2: To what extent do teachers adhere to their allocated approaches?

IPE RQ3: Why and how do teachers use Aila?

IPE RQ4: What is the perceived impact of using Aila?

IPE RQ4.1: What is the impact of using Aila on teachers' self-reported perceptions of their workload?

IPE RQ5: How do teachers' experience and confidence moderate Aila use and any associated changes in workload?

¹https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/lesson_planning_using_ai_lesson_assistant_aila_-_study_plan_-_v.1.0.0.pdf?v=1777504961

Together, these research questions were designed to provide robust evidence on both the impact of Aila on teachers' lesson planning workload and assessing lesson resource quality, and the conditions under which the tool is most effectively used in practice.

Ethics and trial registration

The trial was designed, conducted, and reported to CONSORT standards² and registered on the Open Science Framework (OSF) website of the Center for Open Science.³ The trial record will be updated following publication of the evaluation outcomes.

The trial was conducted in line with [NFER's Code of Practice](#), which sets out principles for ethical research conduct, including data protection, participant welfare, and the responsible handling of data. Ethical approval for the evaluation was obtained through NFER's internal review processes prior to trial commencement, in December 2024. These procedures are aligned with the standards set by professional bodies such as the British Educational Research Association (BERA), the Social Research Association (SRA), and the U.K. Statistics Authority.

Agreement for school participation in the trial was obtained through a signed memorandum of understanding (MoU) completed by a senior leader or the headteacher in each participating school. The MoU outlined the roles and responsibilities of all parties, including expectations for data collection, choice implementation, and adherence to trial protocol. Schools were also asked to distribute participant information sheets to teachers involved in the trial. Relevant documentation, including the MoU, is provided in Appendix E.

Data protection

NFER was the data controller for this evaluation. Oak National Academy was the data controller for any personal data shared on the Aila platform. After the report is published, teacher personal data including direct identifiers (such as names, job role) and the responses from diaries and teacher surveys will be archived in the EEF archive for future research. At this point the EEF will become the data controller for the archived data.

The legal basis for processing personal data is covered by GDPR Article 6 (1) (f): 'Legitimate interests: the processing is necessary for your (or a third party's) legitimate interests unless there is a good reason to protect the individual's personal data which overrides those legitimate interests.' A legitimate interest assessment has been undertaken. The trial fulfils one of NFER's core business purposes (undertaking research, evaluation, and information activities). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence about the impact of teaching techniques used in the classroom. Research cannot be done without processing personal data but processing does not override the data subject's interests.

NFER and Oak National Academy signed a data sharing agreement that governed the collection and sharing of personal data to obtain Aila platform data from Oak National Academy and for it to be used in trial analysis. This agreement included a description of the nature of the data being collected and how it will be shared, stored, protected, and reported. NFER provided an MoU (Appendix E) to schools explaining the nature of the data being requested of schools and teachers and how it will be collected and processed. A privacy notice was also made available for participating schools and teachers (Appendix F).

As part of the sign-up process, NFER collected expressions of interest (Eoi: Appendix G) that were linked to an information sheet (Appendix H). The Eoi included the names, contact details, role of the individual completing the form, their school, and how they heard about the project. Upon receiving the Eoi, NFER requested the completion of a signed MoU from the headteacher or a member of the senior management team. The MoU asked for the names, contact details, and job role for the signatory and the trial lead.

² <https://www.bmj.com/content/389/bmj-2024-081123>

³ <https://osf.io/fwp5z/overview>

Further personal data about teachers was collected throughout the trial via teacher data templates, weekly diary entries, online surveys, interviews, and case study visits. Personal data collected through these activities included names, contact details, job role, length of time teaching, year group taught, subject specialism, if any, current use of Aila or GenAI tools, prior and current usage of Oak National Academy resources, confidence in subject knowledge, pedagogy and technology, use and perceptions of other GenAI tools, perceptions of facilitators, barriers and challenges as well as perceived costs in using Aila, teacher use of Aila and evaluations of Aila process, and perceived impact of using Aila on themselves and on pupil experiences.

Project team

Table 4 details the project teams at NFER, the EEF, and Oak National Academy. These teams were further supported by a Study Advisory Board (SAB). The full membership details are presented in Table 5.

Table 4: Key members of the study team

NFER	
Helen Poet	Project director—responsible for overall delivery of the trial.
Palak Roy	Trial manager—day-to-day management of the trial, delivery of the trial design, and impact evaluation lead.
Katherine Aston	IPE lead—design and delivery of the IPE.
Ruth Staunton	Trial statistician—lead quantitative analysis for the trial.
David Thomas	Lesson resources panel advisor—oversee RQ2 on lesson resource quality.
Daniel Jackson	IPE researcher—undertake IPE case studies and support lesson resource quality.
Kathryn Hurd	Research operations lead—overall data collection and setting communications strategy.
Jishi Jose	Project manager—day-to-day operations for the trial including coordinating data collection and point of contact for the schools.
Max Falinski	Project manager—day-to-day operations for the trial including coordinating data collection and point of contact for the schools.
EEF	
Faizaan Sami	Evaluation manager.
Fabiola Clemente	Programme manager.
Oak National Academy	
Hannah-Beth Clark	AI Quality and Safety Principal.
Reka Budai	Senior User Researcher.
Laura Benton	User Researcher.
Will Gayne	Director of Product Strategy and Impact.
Ewan Wakeman	Senior Data Analyst .
Rachel Urquhart	Interim AI Quality and Safety Principal.

Table 5: The Study Advisory Board (SAB) appointed by the EEF

Name	Organisation	SAB Specialism
Libby Hills	CEO at EdTechnical	Edtech subject expertise; practitioner experience.
Chris Goodall	Bourne Education Trust	Using GenAI in teaching expert.
Ria Drabble	Broadfield Primary School	Primary school teaching experience.
Michael Adkins	University of Nottingham	Methodological expertise.
Jake Anders	University College London	Methodological expertise.

Methods

Trial design

Table 6: Trial design

Trial design, including number of arms		Two-arm, cluster-randomised controlled trial.
Unit of randomisation		School.
Stratification variable(s) (if applicable)		Number of teachers who indicate no/low prior use of GenAI at baseline survey.
Primary outcome	Variable	RQ1: Average time spent per week on lesson planning for KS2 lessons during the ten-week trial period measured by a weekly teacher diary.
	Measure (instrument, scale, source)	Researcher-designed weekly teacher diary.
Secondary outcome(s)	Variable(s)	RQ2: Quality of lesson plans and resources used during the ten-week trial period measured via the average scores given by an independent panel of teachers.
	Measure(s) (instrument, scale, source)	Average score for the lesson and resource material, given by the independent panel of teacher-reviewers, score range: 0–254.
Baseline for primary outcome	Variable	Total expected weekly hours spent in lesson planning, collected before randomisation.
	Measure (instrument, scale, source)	Researcher-designed teacher survey prior to randomisation.
Baseline for secondary outcome(s)	Variable	RQ2 did not have a baseline measure.
	Measure (instrument, scale, source)	N/A

This was a cluster-randomised controlled Teacher Choices trial with randomisation at school level and teachers as the experimental and analysis unit. Schools were randomly allocated to one of two groups: the Aila group or the non-Aila group. The implementation of choices lasted for ten weeks during the autumn term of 2025 and focused on lesson planning carried out by teachers for any subject taught in KS2 (between 27 September and 5 December 2025) including any planning that took place during the autumn half-term break.

Aila group

Teachers in schools allocated to the Aila group were encouraged to use Aila to support the lesson planning for Years 3 to 6 as much as they wanted over a ten-week period during the autumn term 2025. Teachers were able to use other resources as they wished—except for accessing the Oak National Academy website for resources (see below).

Non-Aila group

Teachers from schools in the non-Aila group were asked to continue planning lessons in their usual way over the same ten-week period during the autumn term 2025. They were encouraged not to access the Oak National Academy website—

⁴ This is a deviation from the study plan, where the expectation was that each lesson would be scored out of 20 using a 0–5 scale across four areas of lesson planning.

including Aila—for any lesson planning during this period. They were able to use any other sources that they routinely choose as part of their typical approach to lesson planning, which could have included GenAI tools such as ChatGPT.

Use of Oak National Academy resources during the implementation period

To help isolate the effect of Aila, teachers in both randomised groups were asked not to access the Oak National Academy website during the implementation period. However, teachers were able to choose to re-use any website resources they had downloaded prior to July 2025.

Recruitment and administration

Once schools signed up to the trial, they nominated participating teachers by providing NFER with teachers' names and contact details. In September 2025, teachers were asked to respond to a baseline survey (see Appendix I). Questions in this survey established eligibility for the trial and also formed the baseline measures for the primary outcome and therefore completing the survey was a requirement to take part in the trial. After the baseline surveys were completed (the online survey closed on 22 September 2025), schools were randomly allocated to one of the two groups on 23 September 2025. Participating schools and teachers were notified about their group allocation. Along with this, they were sent a brief trial guidance document produced by the research team. The guidance document provided an overview of their allocated approach, including brief do's and don'ts for the approach. The Aila group guidance document asked teachers to register on the Aila platform and use Aila for their KS2 lesson planning for ten weeks in the autumn term 2025. It also included links to a walk-through video and a brief how-to guide produced by Oak National Academy. Teachers in the non-Aila group were asked specifically not to use Aila when planning KS2 lessons for ten weeks in the autumn term 2025. The trial guidance documents reminded teachers not to access new Oak National Academy resources during the trial, while allowing use of previously downloaded Oak National Academy materials and other existing planning resources. The guidance documents can be found in Appendix D.

Data collection activities for both randomised groups included:

- asking all teachers—in each of the ten weeks of delivery—to complete a short online diary entry regarding the time spent planning or preparing for lessons (in any subject) in that week; this formed the primary outcome (Aila and non-Aila groups);
- conducting case study visits to five primary schools, which included a teacher interview and observation of an Aila-planned lesson (Aila group only);
- sixteen teachers taking part in online interviews;
- all teachers in both groups being asked to submit lesson resources for two lessons that they delivered during the ten weeks (Aila and non-Aila groups);
- asking all teachers—at the end of the ten-week period—to complete a short endpoint teacher survey (Aila and non-Aila groups); and
- inviting all teachers to complete a short online survey in March 2026 (Aila and non-Aila groups).

Participating schools received financial incentives where teachers completed evaluation activities. Teachers needed to complete at least one weekly diary to be included in the primary analysis. Teachers who completed the diary received an incentive of £10 per completed week for their school, up to a total of £100 for completing all ten diary entries. In addition to this, schools also received up to £30 per teacher where teachers submitted lesson resources. Schools participating in case study visits received £100 per visit. Schools where teachers participated in an online interview received a thank-you payment—£50 for those in the Aila group and £30 for those in the non-Aila group (to reflect the different lengths of the interviews). Completion of evaluation activities meant that schools received £400 on average.

The trial focused on measuring how much time teachers spent planning lessons and assessing the quality of their lesson materials rather than pupil outcomes. This was because pupil outcomes were considered longer term as they are theorised to be dependent on consistent teacher usage of Aila (and therefore consistent pupil exposure). This focus was agreed upon during the set-up phase in discussions with Oak National Academy and the EEF. Therefore, the trial's *primary outcome was*

teacher workload, specifically the time spent on lesson planning for participating KS2 teachers over the ten-week implementation period. The *secondary outcome was the quality of lesson and resource materials* used in lessons delivered in the same period.

While the impact evaluation looked at teachers' workload in lesson planning as well as the quality of the materials, the IPE characterised teachers' capability, opportunity, and motivation (see the 'behaviour change framework', Michie, van Stralen and West, 2011), their use of Aila, and the perceived impact for teachers and pupils. This added context to our impact analyses, enabling us to interpret positive, null, or negative effects in relation to teacher practice.

See the Outcomes section for further details.

Participant selection

To participate, schools and teachers needed to meet the following criteria.

School eligibility

- any state-funded mainstream schools in England with (any of) Years 3 to 6;
- not currently encouraging use of Aila;
- willing to not use Oak National Academy website during the trial period;
- has at least one Year 3, 4, 5, or 6 teacher who will be teaching and involved in planning lessons in the autumn term 2025; and
- not participating in the EEF Writing Roots trial.

Teacher eligibility

- teachers who have teaching and lesson planning responsibilities during the autumn term 2025 for any subjects in Years 3, 4, 5, or 6 and are willing to be part of the trial, including those who share class teaching responsibilities (for example, job-share or part-time teachers);
- teachers who plan to create at least some lesson resources from scratch in any subject in Years 3,4,5, or 6 during the trial period and are not intending to use pre-prepared, paid for, or well-established curriculum resources for all of their teaching;
- teachers who are not currently using Aila for lesson planning: teachers who are already using Aila to plan lessons at least once a half term are not eligible to take part; the use of other AI (for example, ChatGPT) is acceptable; and
- teachers who complete the online baseline survey in September 2025.

During the trial teachers were allowed to use existing resources (including Oak National Academy resources, if they were already being used) but they were asked not to use Oak National Academy website to generate new resources during the autumn term.

Recruitment

NFER was responsible for school recruitment for this trial. With the anticipated trial dropout, the power calculations at design stage suggested the target numbers to be recruited for randomisation were 426 teachers from 133 schools (see the Sample Size section for further details). Recruitment activities took place between 15 May and 18 July 2025. NFER approached 15,755 primary schools in England with an EoI form. In addition to this, the EEF and NFER promoted the trial on their respective websites and social media platforms. Interested teachers and leaders expressed initial interest by filling out an online EoI form which was linked to the school information sheet (Appendices G and H). NFER contacted the individuals

at 247⁵ schools who had signed the EoI, where their schools met the eligibility criteria and asked for the headteacher or the senior leader at their school to complete the MoU. In total, 140 schools signed the MoU.⁶ In July 2025, these schools were asked to submit information about participating teachers; this included names, role, contact details, year group, and class name that they will teach in autumn 2025: 109 schools submitted data for 549 teachers. Once the teacher data was received, NFER sent a link to online baseline surveys to all nominated teachers in September 2025. The baseline survey completion was a requirement for individual teacher participation and hence a criterion for a school to be part of the trial. The baseline surveys were completed between 3 and 22 September 2025: 464 eligible teachers from 108 schools completed baseline surveys.

Outcome measures

Baseline measures

The baseline measure for the primary outcome is time teachers expected to spend each week on lesson planning during the trial period, collected prior to randomisation. This was measured through a workload question in the baseline survey where teachers indicated the number of hours they expect to spend per week on lesson planning for KS2 year group(s) they are teaching during the trial period. This measure includes all activities conducted by teachers towards lesson and resource preparation outside of regular class hours, including weekends and evenings, while excluding their time spent on teaching, marking, or administrative tasks. This survey question and the measure was similar (though not identical) to the one used in the ChatGPT Teacher Choices trial (Roy et al., 2024). The question was originally based on questions used in previous workload studies (OECD, 2018; Walker, Worth and Van den Brande, 2019) in which teachers were asked about a wider range of workload domains (such as marking and administration), however, due to our particular focus on lesson planning we only asked about that domain. The spring term timing of the baseline survey in the previous ChatGPT Teacher Choices trial allowed the baseline workload question to focus on a specific and typical pre-trial week. However, in this trial, the baseline survey was completed in the first few weeks of the autumn term and therefore asking about the week prior to survey completion would have increased the risk that it would not be a typical planning week. Consequently, the baseline workload question was updated to ask teachers about their expected weekly hours of planning during the trial period. We asked that they use their experience from the previous year as a guide, if appropriate.

The timing of the baseline survey was at the very beginning of the autumn term and therefore the school year in order to have enough time for ten weeks of the diary and the endpoint survey all within a single term. However, this meant that the workload captured in the baseline survey may not be 'typical' as there had been so little of the school year by that point. We have considered this and we do not anticipate that this will meaningfully reduce the baseline measure's potential to explain variability in the outcome measures, since both are still measures of workload, the baseline was close in time to the trial period (just before randomisation), and each teacher's individual baseline and outcome data has been linked in the analysis models.

There is no baseline measure for RQ2.

Primary outcome

The primary outcome for this trial is teachers' workload in lesson planning over the ten-week trial period, including any planning that took place during the autumn half-term break. This is also the outcome measure for RQ3, RQ4, RQ5, and RQ6. It was measured via a workload question in the weekly teacher diary (Appendix J). The question was repeated every week of the ten-week trial period. This resulted in up to ten data points per teacher of the weekly hours dedicated to lesson planning over the ten-week implementation period during the autumn term 2025. This measure includes all activities conducted by teachers towards lesson planning outside of regular class hours, including weekends and evenings, while excluding their time spent on teaching, marking, or administrative tasks. This could be planning for lessons that are taught in the same week, or for lessons to be taught further in the future.

⁵ Originally NFER had contacted 248 schools but one school wishing to continue using the Oak National Academy website during the trial was deemed ineligible and so was removed from the pool.

⁶ In total, 141 schools signed an MoU but one school withdrew before NFER approached it to collect teacher data.

Teachers are included in the primary analysis if they completed at least one weekly diary. This threshold differs from the approach used in the primary analysis of previous Teacher Choices trial (Roy et al., 2024), where teachers were required to complete 60% of diaries in the primary trial period. The threshold was implemented in that trial because the outcome measure in the previous study was the total hours across a five-week period and when mean imputation was applied to missing weeks before summing, it was limited to a maximum of 40% of data being imputed. In this trial, weeks are treated as replicates—that is, we have up to ten weekly data points per teacher and no summing is applied—meaning that it was not necessary to impute diary data that was not returned. In addition to being unnecessary, imputing diary data for missing weeks would falsely reduce the week-to-week variance. Any diary entries returned have therefore been included in the primary analysis.

Prior to analysis, we checked for the presence of ‘extreme’ values given in the diary question. Responses to numeric questions in the TALIS workload survey (OECD, 2018) are excluded from analysis if they are deemed implausible. For the workload questions on which we model our primary outcome, ‘implausible’ is defined by TALIS as values larger than 120 hours, which are excluded. Collection methodology in our study ensured that lesson planning workload values input for each week could not be below zero or above 50 hours so mirroring the TALIS method to exclude implausible values is unnecessary. The frequency of values outside of the range defined by the mean $\pm$ 3.29*standard deviation (Tabachnick and Fidell, 2013) is reported for both randomised groups in Appendix B. However, we have retained all data for analysis as we have no evidence that they are not true data points.

Secondary outcomes

The secondary outcome (for RQ2) is the quality of lesson plans and resources created during the ten-week implementation period. An independent expert panel of six teacher-reviewers assessed the quality of lesson and resource materials using a quantitative scoring approach (see the Secondary Analysis section for further details).⁷ From the teachers who submitted at least one lesson, we randomly sampled a total of 42—21 from each randomised group—and then randomly selected one lesson from each teacher. Scores from all 42 lessons were included in the analysis for the secondary outcome.

The outcome measure is the score assigned to each lesson by the individual reviewers. Each lesson was scored out of 25 using a 0–5 scale across five key areas of lesson planning.⁸ Therefore, the outcome measure is a numerical variable with a range 0–25, with each lesson receiving separate scores—one from each reviewer. For further details on the sample and scoring process, see the Secondary Analysis section.

Sample size

Please note that the outcomes for this study are all teacher focused rather than pupil focused so there is no separate sample size calculation for FSM pupils. Sample size calculations at design, randomisation, and analysis stages are shown in Table 10 in the Impact Evaluation Results.

Sample sizes at the design stage of this trial were calculated primarily using data from the recently published EEF-funded trial to understand the impact of ChatGPT use on lesson planning (Roy et al., 2024). The ChatGPT trial participants were KS3 science teachers, which is dissimilar to the KS2 teachers participating in this trial, but the primary outcome variable of diary-collected hours spent in lesson planning is similar between the two trials and there are many similarities between the teacher choices—both are GenAI tools that can be used for lesson planning, though ChatGPT can be used for many other purposes too—therefore, we considered the variables sourced from the ChatGPT trial to be appropriate for these sample size calculations.

The primary research question in the ChatGPT trial produced a significant effect size of 0.38. When designing this trial we anticipated that the effect would be smaller due to the increase in prevalence of teachers using GenAI for lesson planning within the sector since the ChatGPT trial, which we expected would result in more GenAI use in our population at baseline and in the non-Aila group during the trial period, potentially reducing the difference between the groups. We therefore

⁷ We chose a scoring process for this project rather than a ranking process (used in the ChatGPT trial) for logistical reasons, including the feasibility of conducting a larger scale ranking exercise during the project timeline.

⁸ Please note this is a deviation from the study plan, where the expectation was that each lesson would be scored out of 20 using a 0–5 scale across *four* areas of lesson planning.

initially (see later in this section for changes) selected a target MDES of 0.3 to ensure we had sufficient power to detect this anticipated smaller difference. Using the pooled standard deviation from the ChatGPT trial, this equates to a fold change of 1.34. The design stage ICC (0.04) and within- and between-school R² parameters (0.06; 0.40) are also sourced from the ChatGPT trial, along with the estimates we employed for attrition between randomisation and analysis—15% teacher attrition and 5% school attrition.

We also reviewed whether TALIS data could inform the target MDES in these power calculations. According to TALIS, primary school teachers in England spend an average of 5.1 hours per week on individual planning or preparation of lessons. Analysis of the time spent on lesson preparation by primary teachers crossed with the question about lesson preparation being a source of stress informed a target MDES of 0.61 (difference in hours of 2.3 hours between the groups). While this would be a meaningful effect, it is unlikely to be the minimum meaningful effect. Our aim was to be able to detect smaller, yet meaningful, differences in workload which we would probably miss if we targeted a MDES of 0.61. We considered the target difference in hours for the MDES of 0.3 of 1.3 hours would still be a meaningful reduction in workload, and a more realistic effect, as outlined in the previous paragraph.

In primary schools in England, the average classroom teacher headcount is 12.43 (this trial allowed for part time teacher participation, so headcount is used rather than full time equivalent). We assumed an equal number of teachers in each primary school year group and assuming 26% signed up among the teachers in the four year groups eligible for this trial (based on our previous experience of recruiting teachers into trials), we estimated an average of 3.2 teachers per school at randomisation.

The parameters detailed in this section so far were used prior to the start of recruitment in a sample size calculation that calculated that 344 teachers from 126 schools would need to be analysed to give an MDES of 0.299. With the anticipated trial dropout, the target numbers to be recruited for randomisation were 426 teachers from 133 schools. This sample size calculation assumed a two-level random structure as used in the ChatGPT trial (teachers within schools). However, during the study plan peer review process, reviewer feedback highlighted that using a three-level model (weeks within teachers within schools) for the sample size calculation would be more appropriate to the planned analysis and should improve the statistical power. Unfortunately, due to the rapid timelines that meant this study plan had to be drafted in parallel with recruitment, this insight was gained after the majority of recruitment had already taken place. Therefore, the design stage numbers detailed in Table 10 below use the three-level model ICC, R², and number of weeks per teacher parameters (from the ChatGPT trial data but not published) and maintain the target number of schools (133) and teachers (426) from the two-level sample size calculation. The achieved analysis MDES in this scenario is 0.239. The minimum detectable fold change (MDFC) is estimated as $10^{(MDES \times \text{Standard Deviation from ChatGPT Trial})}$ where the standard deviation in the ChatGPT trial (on the log₁₀ scale) was 0.424.

Numbers in the randomisation stage column of the table are mostly the same parameters from the design stage, with the exception of the total number of schools and teachers, and therefore the average cluster size. This was updated to reflect the numbers randomised (464 teachers from 108 schools). Although the number of schools recruited is lower than the target (108 vs 133), the number of teachers is higher than the target (464 vs 426) so the MDES at randomisation stage is very similar to the MDES at design.

All parameters in the analysis stage column of the table have been updated to be the numbers from the analysis of this trial. All sample size calculations were conducted using the R statistical software and the package PowerUPR (Buluş et al., 2021).

Randomisation

This is a cluster randomised trial with randomisation at school level. We elected to randomise at school level rather than teacher level due to the risk of contamination when teachers within a school collaborate on lesson planning activities. This increased risk was judged to outweigh the advantage of smaller sample size that would be associated with teacher-level randomisation. All sample size calculations and all statistical analyses account for the clustering of teachers within schools.

Schools were randomly assigned to one of two arms with equal allocations (1:1). We stratified the randomisation by the number of teachers in a school who indicate low or no prior use of GenAI in lesson planning in the baseline survey. This

variable was a categorical grouped variable with five groups—0, 1, 2 to 3, 4 to 5, and 6+. This variable was chosen as a stratifier to improve the likelihood of balance among groups in the subgroup analysis. Table 7 presents school numbers in each stratum and randomised group.

Recruitment was carried out by NFER’s Research Operations team who were also liaising with schools and teachers on a regular basis throughout the trial. Randomisation was performed by the trial statistician on 22 September 2025. It was carried out using trial specific school IDs (not potential identifiers such as URN or school name) and allocation groups were generically labelled (group A, group B) until the randomisation was complete and quality assurance (by a statistician outside of the project team) had been signed off. At this point, the trial groups were randomly assigned to the generic labels. R code used for randomisation is included in Appendix K for transparency.

Table 7: Number of schools randomly allocated to each treatment arm by prior Gen AI use

Number of ‘low or no prior Gen AI’ teachers	Aila	Non-Aila	Total
0	12	13	25
1	16	16	32
2 to 3	9	8	17
4 to 5	8	8	16
6 or more	9	9	18
Total	54	54	108

Statistical analysis

Primary analysis

RQ1: What is the impact of Aila on the average time spent per week on lesson planning for KS2 lessons during the ten-week trial period?

The primary research question was analysed on an intention-to-treat basis. We used a linear mixed effects model to estimate the effect of being in the Aila group on teacher lesson planning time in the ten-week trial period, controlling for baseline lesson planning time and stratification group by means of covariates and accounting for clustering of teachers within schools and weeks within teachers. The regression model was defined by:

$$Workload_{ijk} = \beta_{0k} + \beta_{0jk} + \beta_1 Aila_k + \beta_2 BaseWorkload_{jk} + \beta_3 StratificationGroup_k + \epsilon_{ijk}$$

where:

- $Workload_{ijk}$ is the KS2 lesson planning workload in hours for teacher j in school k for week i ; this was measured using the weekly teacher diary;
- $BaseWorkload_{jk}$ is the expected weekly KS2 lesson planning workload in hours estimated by the teacher prior to randomisation for teacher j in school k as measured by the baseline teacher survey;
- $Aila_k$ is a variable indicating whether school k was randomly assigned to use Aila or not;
- $StratificationGroup_k$ is a variable indicating which stratification group school k belongs to;
- β_{0k} is the intercept of school k (modelled as a random effect);
- β_{0jk} is the intercept of teacher j in school k (modelled as a random effect);

- β_1 is the coefficient of interest estimating how much difference being assigned to the Aila group makes to teacher workload;
- β_2 is a coefficient estimating the association between baseline teacher workload and workload in the ten-week trial period;
- β_3 is a coefficient estimating the association between stratification group and workload in the ten-week trial period; and
- ε_{ijk} is the residual error term for teacher j in school k in week i .

Distributional assumptions of the model were assessed through visual inspection of the residual plots. A strong heterogeneity of variance was apparent, so a $\log_{10}$ transformation of the outcome variable was applied. More detail is reported in the Impact Evaluation Results section.

Secondary analysis

RQ2: How does the quality of lesson plans and resources in the Aila group compare to the quality of lesson plans and resources in the non-Aila group?

The methodology for addressing this RQ was built on the approach that we used for the ChatGPT trial and is outlined in the study plan. The purpose was to provide data to confirm whether there is any statistical evidence that the quality of Aila produced lessons was at least as good as that of non-Aila lessons. Broadly, this involved the following phases:

Collection of lesson plans and resources

During the trial, teachers who were participating in each arm were asked to provide lesson resources for lessons that had been planned during the trial period. The lesson resources were requested at two points during the process: at weeks three and seven. At each timepoint teachers were asked to provide the materials from their most recent lesson as part of the trial. This was to dissuade teachers from specifically producing lesson materials to submit that may be of a higher standard to their normal planning as part of the trial.

When teachers submitted the lessons, they were requested to complete a 'lesson overview form'. To simplify the process for the teachers this was collected via a MS form with most fields being entered by multiple choice. This collected summary information about the lesson, specifically:

- year group for the lesson—Year 3 to Year 6 or mixed;
- ability level for the class—high, middle, support, mixed;
- notes on class ability—optional open field to cover specifics such as SEND, gifted and talented, groupings used in class;
- KS2 subject—11 subject areas of the KS2 curriculum plus a tick for any other; teachers could select more than one of the subject areas;
- topic of lesson—open text entry for teacher to add;
- learning objective of lesson—open text entry for teacher to add; and
- brief overview of the lesson structure—open text entry for teacher to add.

Teachers were also asked to confirm that the lesson was planned during the study period and provide any information about any GenAI platforms used during the production of the lesson resources (including Aila). However, this information was redacted before being shared with the review panel as they were 'blinded' as to the purpose of the trial.

Lesson resources were included if the teacher provided at least two valid lesson submissions—resources and completed lesson overview form. When entered into the review, only one lesson was used per teacher being randomly selected from those provided. This again is a measure against teachers producing lessons at a different standard to normal for the review process. Overall, 108 teachers provided resources, however nine were excluded from the sample due to missing information

or incomplete resources. We randomly selected 42 teachers (equal allocation from each arm) then randomly selected one lesson submitted by each of those teachers for analysis. Fifteen of the 21 selected Aila group lessons (71%) were noted by the teacher as ‘created using Aila’. This is representative of all submitted lessons by the Aila group teachers, where 68% of lessons were noted as ‘created using Aila’.

Development of a rubric to assess the quality of lesson materials

The methodology for reviewing lessons requires a common understanding from the reviewers of what are the important aspects of lesson quality to consider and score. This approach was piloted in the ChatGPT trial by the development of a lesson scoring rubric, which divided lesson quality into four categories and provided a scoring matrix from zero to five for each. Following on from that project, we further developed the rubric specifically for use with KS2 lessons. The rubric was reviewed internally at NFER before being shared with the EEF and discussed with KS2 practitioners at a SAB meeting. The rubric was amended following the outcomes of this meeting and a second draft was distributed to the EEF, Oak National Academy, and the SAB panel members for further review and feedback. NFER collated all of the feedback from these reviews and made a final decision on the rubric details. The final version was also reviewed for clarity by other NFER researchers who had recently been KS2 teachers. The broad amendments to the rubric following this process are listed below.

The number of lesson review categories was increased from four to five since the study plan.⁹ There were four lesson review categories originally: extent that activities engage students with their learning and check their understanding, appropriateness for the age group and the ability level of the class, and quality and accuracy of the subject content. As a result of feedback from the SAB, the category for ‘engagement with pupils and checking of pupil learning’ was expanded into two new categories: ‘extent that activities involve pupils with their learning progression’ and ‘extent that activities check the understanding of the learners’. This refinement was made to distinguish between the extent to which activities supported pupils’ learning progression and the extent to which they assessed pupils’ understanding, reflecting two conceptually distinct aspects of lesson resource quality.

Additional examples of activities that indicate the higher scoring levels were included and adapted to the language of KS2 teaching

- Specific reference was included to ‘adaptive pedagogy’ as the current term used in teacher training for matching activities to the range of learners in the classroom.
- Rubric categories were linked to the Early Career Framework (ECF),¹⁰ standards which are provided as guidelines for good practice for new teachers. A second document was produced which directly linked the rubric categories to the ECF standards and provided specific examples of good practice from the ECF as examples of things to look for in the lesson planning resources for the specific categories.
- The rubric document and ECF standards document were shared with the reviewers and are provided in Appendices L and M.

Recruitment and training of reviewers

Reviewers for this process were recruited from NFER’s pool of associate teachers. A recruitment notice was sent out asking for current teachers to review KS2 lesson plans. No mention was made of the nature of the study. It is important for the review process that the teachers involved in the review are ‘blinded’ to the specific purpose of the trial. Therefore, the recruitment notice specified only the need to review KS2 lesson plans and resources: no mention was made of AI involvement.

The eligibility criteria to join the review panel were:

- a qualified teacher who has completed their early career induction period;

⁹ The rubric was being refined and updated as a result of the feedback received from the SAB, which was completed after the study plan was published.

¹⁰ [Early Career Framework](#)

- currently working in a primary school as a classroom teacher (excluding supply or PPA cover teachers); and
- a KS2 teacher.

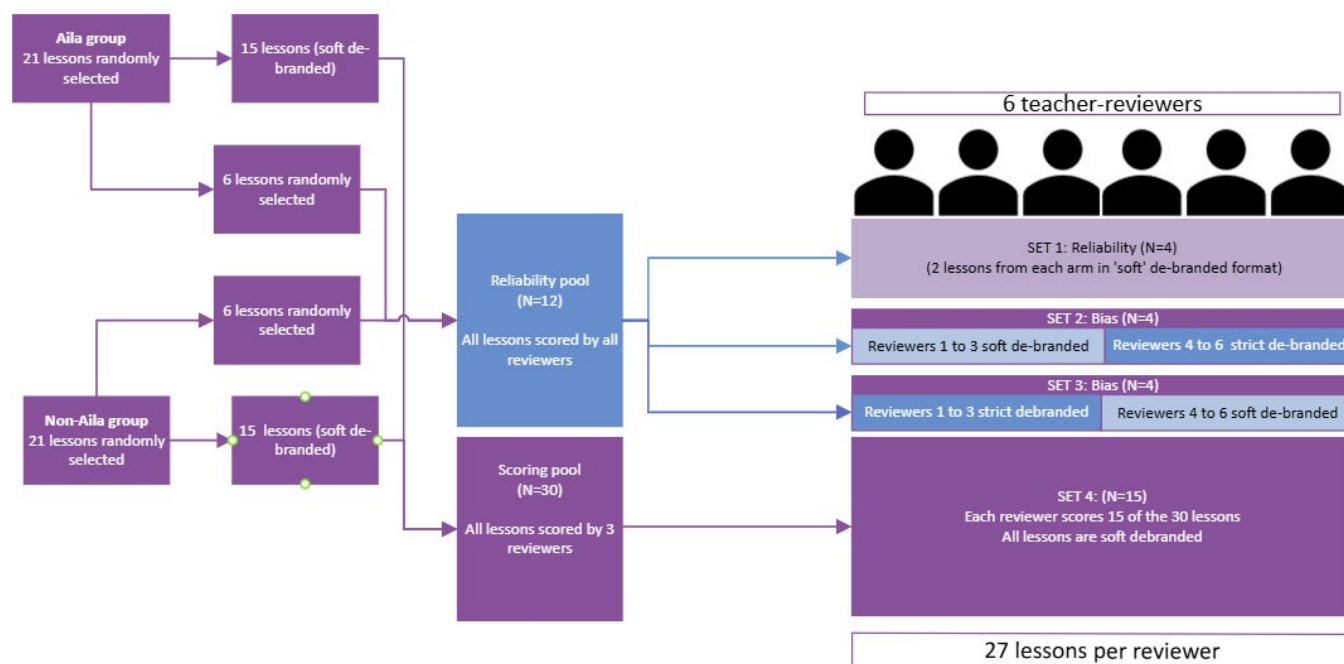
It was specified that we were looking for teachers with a range of experience and different subject specialisms and this would be taken into account when choosing the panel.

From the initial applicants, a shortlist was developed based on the above criteria plus availability during the scoring period. Teachers from schools that participated in the trial (either arm) were excluded from selection. The final selection from the shortlist was conducted after the lesson resources had been submitted and collated by NFER so that the expertise of the panel would match the approximate distribution of subjects received in the trial. The panel composition ended up as two English/literacy specialists, one mathematics/numeracy specialist, one RE specialist, and two generalists (covering science/geography/history and SEND specialism). Panel members were paid in line with similar work undertaken by NFER associates, reflecting the level of commitment and time required.

Allocation of resources to reviewers

In total, sufficient lesson plans and materials were collected for the full programme of review of lesson resources to be reviewed (Figure 2).

Figure 2: Distribution of lesson resources between the reviewers



The complete set of lesson resources for the 21 lessons from each arm (42 in total) were collated into folders along with the lesson review forms. The folders were named with a random four-letter code. All of the materials were 'soft de-branded'. This meant that all references to any external scheme of work, software used to produce the materials, school or MAT name, or any names of member of staff or children were removed. This was to maintain confidentiality and 'blind' the reviewers to the source of the lesson materials. However, the formatting, use of fonts, and colour schemes were maintained so that lesson resources could be reviewed as close as possible to their format of delivery.

This model included a measure of marker reliability by inclusion of a proportion of lessons scored by all six reviewers. Based on the model shown in Figure 2, six lessons from each arm were randomly allocated into a reliability pool of lessons. The first set of four responses, two from each arm, were scored by all of the reviewers immediately after the training and some feedback provided on areas of focus with the review. All of the lessons selected for review as part of sets two and three were also 'strict de-branded', which meant that in addition to the soft de-branding, the colour scheme was removed (presented

as greyscale) and font styles were changed to Arial font, where possible. Each reviewer reviewed all of lessons in these sets but saw one set in soft de-brand and the other in strict de-brand. The purpose of this was to identify if there were any effects of strict de-branding on the perception of the lesson quality.

The remainder of the lessons were split between all of the reviewers as shown in Figure 2; the 30 lessons in set four were distributed individually to the reviewers. In a deviation from the original plan these lessons were distributed individually rather than in sets so that we were able to more evenly distribute them such that every lesson was reviewed by at least three of the reviewers and that each reviewer scored an equal number of lessons in common with each of the other reviewers. In total, across all sets, each reviewer completed 27 reviews (comprising 13 or 14 lessons from each arm of the study). However, as stated, the reviewers were 'blinded' to the purpose of the study. All of the lessons reviewed by each reviewer (except for reliability set one) were mixed into a single pool for each reviewer and listed in the alphabetical order of the random four-letter ID code. Therefore, reviewers were blinded to the different arms of the study.

The review process

Training was held to familiarise the reviewers with the scoring rubric, the NFER systems, and clarify the process required. The training meeting was held as a remote meeting on 17 January 2026. Prior to this, the reviewers had been provided with the rubric and the Early Careers Framework mapping document. The training reviewed the contents of these documents and how they can be used to score lesson materials. The NFER researcher conducting the session then worked through two sets of exemplar materials demonstrating how to apply the rubric to the materials. The reviewers were then asked to independently apply the rubric to a further sample of four lessons before reconvening to share outcomes and discuss the shared understanding of the process.

Following the training, all of the reviewers were supplied with the four lessons from set one and asked to score these independently. These scores were then reviewed by the NFER researcher and generalised feedback provided. Once the reviewers had acknowledged the feedback, they then completed the reviews for all of the remaining lessons. All scores were entered separately for each of the five categories of lesson quality for each lesson and the reviewer was able to provide written feedback on any aspect of the lesson. All reviewers proceeded independently with their allocation unable to see the scores or allocation of the other reviewers. The reviewing period proceeded until 31 January 2026. Following the completion of the reviewing the spreadsheets were downloaded and the analysis completed.

Analysis

As detailed in the study plan, inter-rater reliability (percentage agreement) was calculated for the reliability pool set one resources using the formula:

$$IRR = \frac{\sum_{i=1}^{Ncomp} ScoresAgree}{Ncomp}$$

where *Ncomp* is the number of pairwise comparisons between raters across all set one resources and *ScoresAgree* is a binary (0/1) variable indicating whether the scores of the two raters for a given resource are the same (1) or different (0). We have reported the IRR for all resources in set one (*Ncomp* = 60) and separately for the Aila group and non-Aila group resources in set one (*Ncomp* = 30 for each).

This metric outlined in the study plan looks for exact agreement on scores between reviewers. However, as the rubric had a continuous scale (0 to 25) perfect agreement between reviewers is not to be expected in this case and so we subsequently felt that this strict metric did not feel appropriate. Therefore, in addition to the percentage agreement detailed above, several additional inter-rater reliability metrics were calculated. First, the analysis population was expanded to all resources, not just the set one resources that had been seen by all six reviewers. Second, the percentage agreement with a tolerance of ± 2 was calculated for all resources and separately by randomisation group. Finally, the intraclass correlation coefficient (ICC) was calculated for all resources and separately by randomisation group. These calculations were not specified in the study plan: the decision to carry out and report them was made post-hoc.

To check for any bias associated with the lesson resource 'branding', the scores from the reviews of the reliability pool set two and set three resources were fitted as the outcome in a linear mixed effects model with randomisation group (Aila/non-

Aila), branding (soft de-branded/strict de-branded), and their interaction as the fixed effects and reviewer as the random effect. Means and confidence intervals are reported for all four combinations of the fixed effects along with the p -value from a likelihood ratio test testing for the inclusion of the interaction term.

The scores from each reviewer for the soft de-branded scoring pool resources, and for the ‘soft de-branded’ reliability pool resources, were used as the outcome in the analysis model to answer RQ2. This was a linear mixed effects model with randomisation group as the fixed effect and reviewer as the random effect. The randomisation group coefficient—the difference between the groups—was compared to a non-inferiority margin of -2.5 to determine if the quality of lesson-plans in the Aila group is not inferior to the quality of lesson-plans in the non-Aila group. Note that this is a deviation from the study plan where the intended limit was -2. The limit was selected at 10% of the scale range as we believed it represented the smallest difference that would be meaningful. Between study plan publication and running the scoring sessions with the expert panel, the number of scoring categories was increased from four to five, each still with a scale of 0 to 5 (see Development of a Rubric... section above). Therefore, the range of the total scale increased to 25 and we decided to maintain the limit at 10% so it was increased from 2 to 2.5.

This analysis to answer RQ2 is undertaken using an ‘intention-to-treat’ framework. As noted in the study plan, if more than 25% of the selected Aila group resources were not created using Aila, two further analyses would be undertaken to investigate if dilution was introduced by the lack of Aila usage in the Aila group. As described above, 71% of the 21 selected Aila group resources were noted as ‘created using Aila’ so these further analyses were undertaken and are reported as follows

- First, the subgroup of Aila-created resources in the Aila group were compared to the non-Aila resources from the non-Aila group (all 21 in this selection) where weighting was applied to the non-Aila group to ensure that the groups were matched on various teacher and lesson characteristics (teacher characteristics: number of years teaching experience, workload from the baseline survey, workload as a source of stress from baseline survey; lesson characteristics: target year group, subject, target class ability—that is, mixed ability, high ability or support class).
- The second additional analysis used an instrumental variable approach using two-stage, least squares methods to estimate the effect on lesson quality of using Aila to plan the lesson. For the first stage, the ‘Aila-created lesson’ indicator was regressed on random assignment; for the second stage, lesson quality was regressed on each lesson’s predicted ‘Aila-created lesson’ value from the first stage. The results from both stages are reported.

Sensitivity analyses

The purpose of both the sensitivity analyses was to compare the magnitude, direction, and significance of the intervention group coefficient in these models with the same coefficient in the primary analysis. The discussion on the findings is around the reliability of the primary conclusion and/or hypothesis generation for future research.

RQ3: What is the impact of Aila on average time spent per week on lesson planning for KS2 lessons excluding the first week of Aila use? [to allow for familiarisation]

To answer this research question, we repeated the analysis detailed for RQ1 but, for each teacher in the Aila group, excluding the first week of the ten-week trial period where they indicate any Aila use in their weekly diary.

RQ4: What is the impact of Aila on average time spent per week on lesson planning for KS2 lessons excluding half term week(s) [to allow for different amounts of, and variability in, planning during holiday time]

Since the RQ1 analysis includes teacher workload during the half-term break—and not all teachers may have planned lessons during that time, or some may have used it as an opportunity to plan more—we also carried out a sensitivity analysis that excludes the autumn half-term break for each school to allow for more consistent teacher workload across the entire sample.

To answer this research question, we repeated the analysis detailed for RQ1 but excluding the autumn half-term break for each school. The week(s) excluded differ between schools; this information was identified on the website of every participating school with the dates confirmed at the individual school level.

Subgroup analyses

The subgroup analysis models do not contain the stratification group variable that is present in the primary analysis model. This variable was removed during the analysis due to collinearity with the ‘prior use of GenAI’ variable. In retrospect, this should have been anticipated during the study plan stage and the inclusion of the stratification group variable in the subgroup analysis specification was an error.

RQ5a: What is the impact of Aila on the average time spent per week on lesson planning for KS2 lessons among the subgroup of teachers who indicate low/no prior use of GenAI in lesson planning?

As the use of GenAI tools by teachers has increased recently, we anticipated that the primary analysis would include teachers using other GenAI tools, which evidence shows can help teachers to save time when planning lessons (Roy et al., 2024). This subgroup analysis was included to be able to explore whether there is an impact of using Aila for those new to GenAI, who we would expect to not already be benefiting from reduced workload.

We performed a subgroup analysis to explore the presence of heterogeneous treatment effects across the ‘prior use of GenAI’ subgroups (response to the prior use baseline survey question). We expanded the primary analysis linear mixed effects model to include an interaction with the subgroup variable and tested for differences between Aila and non-Aila groups within each level of ‘prior use of generative AI’. The teacher-level regression model was defined by:

$$Workload_{ijk} = \beta_{0k} + \beta_{0jk} + \beta_1 Aila_k + \beta_2 PriorGenAI_{jk} + \beta_3 (Aila_k * PriorGenAI_{jk}) + \beta_4 BaseWorkload_{jk} + \epsilon_{ijk}$$

where:

$Workload_{ijk}$ is the KS2 lesson planning workload in hours of teacher j in school k for week i , measured using the weekly teacher diary;

$BaseWorkload_{jk}$ is the expected weekly KS2 lesson planning workload in hours estimated by the teacher prior to randomisation for teacher j in school k as measured by the baseline teacher survey;

$Aila_k$ is a variable indicating whether school k was randomly assigned to use Aila or not;

$PriorGenAI_{jk}$ is a variable indicating which prior use of GenAI subgroup teacher j (in school k) belongs to;

β_{0k} is the intercept of school k (modelled as a random effect);

β_{0jk} is the intercept of teacher j in school k (modelled as a random effect);

β_1 is the coefficient estimating how much difference being assigned to the Aila group makes to teacher workload;

β_2 is a coefficient estimating the association between prior use of GenAI and teacher workload;

β_3 is a coefficient estimating effect of the interaction between Aila group and prior use of GenAI;

β_4 is a coefficient estimating the association between baseline teacher workload and workload in the ten-week trial period; and

ϵ_{ijk} is the residual error term for teacher j in school k in week i .

In addition, we ran separate models for the four levels of ‘prior use of GenAI’. These took the exact form of the primary analysis described for RQ1 but with only the subset of teachers in each prior use of GenAI group and excluding the stratification group variable (‘split sample’ analysis).

RQ5b: What is the impact of Aila on the average time spent per week on lesson planning for KS2 lessons among the subgroup of teachers who indicate low/no prior use of GenAI in lesson planning and who intend to plan a high proportion of their lessons from scratch?

This subgroup analysis was included to be able to explore whether there was an impact of using Aila for those new to GenAI and who are intending to plan a higher proportion of their lessons from scratch.

The baseline survey had five categories for ‘planning from scratch’: 5%–20%, 21–40%, 41–60%, 61–80% and 81–100%. However, due to low numbers of teachers in some of the percentage categories, we combined the top two and combined the bottom two to give three categories: 5%–40%, 41–60%, and 61–100%. We performed a subgroup analysis to explore the presence of heterogeneous treatment effects across the interaction between ‘prior use of GenAI’ and the three newly created ‘from scratch lesson planning’ subgroups (responses to the prior use baseline survey question and the from scratch baseline survey question). We expanded the primary analysis linear mixed effects model to include an interaction with both of the subgroup variables and tested for differences between Aila and non-Aila groups within each combination of ‘prior use of generative AI’ and ‘from scratch lesson planning’ levels. The teacher-level regression model was defined by:

$$\begin{aligned} \text{Workload}_{ijk} = & \beta_{0k} + \beta_{0jk} + \beta_1 \text{Aila}_k + \beta_2 \text{PriorGenAI}_{jk} + \beta_3 \text{FromScratch}_{jk} + \beta_4 (\text{Aila}_k * \text{PriorGenAI}_{jk}) \\ & + \beta_5 (\text{Aila}_k * \text{FromScratch}_{jk}) + \beta_6 (\text{PriorGenAI}_{jk} * \text{FromScratch}_{jk}) + \beta_7 (\text{Aila}_k * \text{PriorGenAI}_{jk} \\ & * \text{FromScratch}_{jk}) + \beta_8 \text{BaseWorkload}_{jk} + \varepsilon_{ijk} \end{aligned}$$

where:

Workload_{ijk} is the KS2 lesson planning workload in hours of teacher j in school k for week i , measured using the weekly teacher diary;

BaseWorkload_{jk} is the expected weekly KS2 lesson planning workload in hours estimated by the teacher prior to randomisation for teacher j in school k as measured by the baseline teacher survey;

Aila_k is a variable indicating whether school k was randomly assigned to use Aila or not;

PriorGenAI_{jk} is a variable indicating which prior use of GenAI subgroup teacher j (in school k) belongs to;

FromScratch_{jk} is a variable indicating which from scratch lesson planning subgroup teacher j (in school k) belongs to;

β_{0k} is the intercept of school k (modelled as a random effect);

β_{0jk} is the intercept of teacher j in school k (modelled as a random effect);

β_1 is the coefficient estimating how much difference being assigned to the Aila group makes to teacher workload;

β_2 is a coefficient estimating the association between prior use of GenAI and teacher workload;

β_3 is a coefficient estimating the association between from scratch lesson planning and teacher workload;

β_4 is a coefficient estimating effect of the interaction between Aila group and prior use of GenAI;

β_5 is a coefficient estimating effect of the interaction between Aila group and from scratch lesson planning;

β_6 is a coefficient estimating effect of the interaction between prior use of GenAI and from scratch lesson planning;

β_7 is a coefficient estimating effect of the three-way interaction between Aila group, prior use of GenAI and from scratch lesson planning;

β_8 is a coefficient estimating the association between baseline teacher workload and workload in the ten-week trial period; and

ε_{ijk} is the residual error term for teacher j in school k in week i .

In addition, we ran separate models for each of the combination of levels from the ‘prior use of GenAI’ by ‘from scratch lesson planning’ subgroups. These took the exact form of the primary analysis described for RQ1, but with only the subset of teachers in each prior use of GenAI/from scratch planning group, and excluding the stratification group variable (‘split sample’ analysis).

Analysis in the presence of non-compliance

RQ6: What is the impact of using Aila on the average time spent per week on lesson planning for KS2 lessons among ‘compliant’ teachers during the ten-week trial period?

In this trial compliance is considered a continuous variable, defined as the percentage of lessons in which Aila was used for lesson planning. This includes any use of Aila, even if only for a small part of the lesson. This is calculated from responses to a question in the weekly diary where respondents are asked to report the percentage between zero to 100. We have assumed two-sided non-compliance as Aila is freely available, so while use of Aila in the non-Aila group was discouraged, it was not impossible.

We conducted an analysis to estimate the complier average causal effect (CACE). An instrumental variable (IV) analysis was performed using two-stage, least squares methods to estimate the effect of teacher’s compliance with their group allocation on lesson planning workload. For the first stage, the compliance indicator was regressed on random assignment together with covariates from the primary analysis model. For the second stage, lesson planning workload was regressed on each teacher’s predicted compliance value from the first stage together with covariates from the primary analysis model. The coefficient for predicted compliance in this second stage is the CACE (complier average causal effect) estimate for the effect of compliance (as defined above) on lesson planning workload. Results from both stages are reported.

Missing data analysis

The diary was set up so that when a teacher clicked their individual link they were shown the earliest incomplete week, so all missing data is at the end of a teacher’s trial participation, that is, if there is missing data for a teacher, that means they stopped participating in the trial since the instrument was otherwise designed to prompt teachers to populate data for each consecutive week of the trial period so that there are no gaps. If the total number of teachers with any missing data is less than 5%, missing data would not typically bias the primary impact estimates, regardless of the pattern of missingness (Schafer, 1999). In this scenario, a complete case analysis of all teachers who complete all ten weeks of diaries would have been undertaken but no further investigation of missing data would have occurred.

Since more than 5% of teachers have any missing data, patterns of missingness have been investigated by fitting a binomial generalised linear mixed effects model with a binary indicator identifying ‘any missing data’ as the response against teacher characteristics collected in the baseline survey—that is, baseline workload, whether trial lead, current role, main teaching subject, number of years of teaching experience, whether trainee/ECT, and ITT subject. School was the random effect. Characteristics which were significantly associated with likelihood of missingness were included as an additional covariate in a repeat of the primary analysis as a sensitivity analysis. The complete case analysis described above was also performed.

Rate of missing data is reported for each week of the study in each of the Aila and non-Aila groups. We ran a chi-squared test to explore whether teachers who have any missing data are unevenly split across the Aila and non-Aila groups. Mixed effects models are utilised in all impact RQs and these models are considered robust to missing data (Laird and Ware, 1982).

Estimation of effect sizes

For research questions that used time spent in lesson planning as the outcome (RQ1, RQ3, RQ4, RQ5, RQ6), differences are reported as fold changes, calculated as the inverse $\log_{10}$ of the β_1 coefficient.¹¹

¹¹ Where the treatment group coefficient would usually (without log transformation of the outcome variable) be the difference between control and treatment arms ($Y_t - Y_c$), the fold change is the ratio (Y_t / Y_c). Fold change values greater than one represent a numerically positive effect, i.e. treatment is greater than control and values less than one represent a numerically negative effect. Significance is declared where confidence intervals do not contain one, i.e. reported p -values are testing the hypothesis that the effect is significantly different to one.

In line with the EEF statistical guidelines (EEF, 2022) we have also reported differences as effect size for the all outcomes, and this is the only reported outcome for RQ2 as the scoring scale for RQ2 is not meaningful to report readers. Effect sizes are reported as a standardised effect size, calculated using the following equation:

$$ES = (\bar{Y}_T - \bar{Y}_C)_{\text{adjusted}} / sd_{\text{pooled}}$$

where

$(\bar{Y}_T - \bar{Y}_C)_{\text{adjusted}}$ denotes the difference in means between trial groups on the $\log_{10}$ scale adjusting for baseline workload. In our models this is equivalent to the coefficient β_1 so this model coefficient was used in the calculation.

sd_{pooled} is the square root of the total variance from the null model (with only intercept as the fixed effect), calculated as

$$s_p = \sqrt{\sigma_S^2 + \sigma_T^2 + \sigma_R^2}$$

where σ_S^2 is the school-level variance, σ_T^2 is the teacher-level variance, and σ_R^2 is the residual variance.

Confidence intervals for each impact estimate have been estimated by adding/subtracting from the point estimate the standard errors of the Aila group coefficient multiplied by the left-tailed inverse of the Z-distribution with a probability of 2.5% and the number of degrees of freedom associated with the size of the sample. The confidence around the impact estimates was converted to effect size confidence intervals using the same formula as the effect sizes themselves.

Estimation of ICC

School-level ICCs are calculated as the ratio of the school-level variance to the total variance (school-level plus teacher-level plus residual). Teacher-level ICCs are calculated as the ratio of the teacher-level variance to the total variance (school-level plus teacher-level plus residual). Both are reported in Tables in Appendix B.

Implementation and process evaluation

Research methods

The IPE characterised teachers' confidence, opportunity, and motivation based on the behaviour change framework (Michie, van Stralen and West, 2011), their use of Aila, and the perceived impact for teachers and pupils. This adds context to our impact analyses and supports interpretation of the impact findings in relation to teacher practice.

IPE RQ1: What is usual practice in lesson planning in KS2 for teachers in the trial? [usual practice]

- To what extent are teachers using GenAI tools (other than Aila) to plan lessons (both groups, prior to, and during, the trial period)?
- What is usual practice for lesson planning in the non-Aila group? What sources do these teachers use in their lesson planning?

We explored usual practice before randomisation for all participating teachers (including those allocated to Aila and non-Aila approaches), including how this varied for different schools, teachers, and curriculum subjects. This included understanding teachers' prior use and perceptions of Oak National Academy resources and Aila as well as any use of other GenAI tools for lesson planning. As Aila supports teachers (a) planning 'from scratch' (that is, developing a lesson plan and finding/creating resources) and (b) creating specific teaching materials (starter/exit quizzes, glossaries, and/or comprehension tasks) to supplement an existing lesson, we explored the prevalence of these approaches to lesson planning. During the trial we also captured what 'usual practice' looks like in the schools allocated to the non-Aila group during the implementation period, and any changes from previous practice.

We have contextualised these findings in reference to other research on lesson planning approaches, both to characterise the group of teachers involved in the trial and to explore any developments in this rapidly evolving area.

IPE RQ2: To what extent do teachers adhere to their allocated approaches? [adherence]

- To what extent do teachers in the Aila group use Aila regularly through the trial? How do patterns of use vary across the ten weeks?
- To what extent do teachers in the non-Aila group use Aila? [non-adherence]
- To what extent do teachers in either group use the Oak National Academy curriculum materials for lesson planning during the trial period?
- What are the reasons for adherence or non-adherence?

For teachers in the Aila group, we identified the proportion who adhered to the approach, defined as any use of Aila within the choice implementation period. This complements the analysis of frequency of Aila use in the Impact Evaluation.

Teachers in the non-Aila group were asked to continue with their usual practice for lesson planning during the trial and not to use Aila. We will describe the proportion of teachers who adhered to not using Aila, and the frequency of use of Aila within this group (note that usual practice is captured in IPE RQ1).

For each teacher group, we explored the reasons for any non-adherence.

IPE RQ3: Why and how do teachers use Aila? [fidelity, teacher responsiveness]

- In which scenarios do teachers choose to use Aila? When do teachers choose not to use Aila?
- To what extent do teachers use the (a) lesson planning and (b) teaching materials functions of Aila?
- To what extent do teachers interact with Aila and use refinements—for example, to meet the needs of their class/pupils, their teaching style, their curriculum?
- To what extent do teachers adapt or modify the outputs produced by Aila to suit their class after they have been downloaded? Is this out of desire or necessity?
- Does this vary by subject?
- Does this vary by teacher characteristics?
- What do teachers like about Aila? Which aspects of Aila do teachers find most useful?
- How long did teachers need before they felt they could use Aila effectively?
- What are the barriers and challenges to using Aila?

We identified what motivates teachers in the Aila group to use Aila and scenarios in which they choose to use it (including its use for different subjects). We described how teachers interact with Aila (including their use of refinements) and how they use Aila-generated resources in finalising and teaching lessons.

Aila emphasises the importance of teacher decision-making in the planning process, supporting the teacher with time-consuming aspects of planning. The Aila process encourages teachers to interact with Aila to guide and refine planning. We recognise that there are no pre-specified metrics for interactions or refinements, and that ‘good’ use will vary by context. Noting that, we explored fidelity by characterising how teachers interact with Aila in the context of their experience and teaching context. This includes their use of interaction, refinements, and how Aila outputs are incorporated in finalised lesson plans and resources.

We explored teacher perceptions of the process of using Aila, the quality of the process, and the quality of Aila-generated resources.

IPE RQ4: What is the perceived impact of using Aila? [perceived impact]

IPE RQ4.1: What is the impact of using Aila on teachers' self-reported perceptions of their workload? [perceived impact]

- If time is saved, how is this time used by teachers?
- What is the perceived impact of using Aila on the quality of the lesson resources produced?
- How does using Aila to plan a lesson influence teachers' approaches to teaching the lesson, if at all?
- How do teachers perceive that pupils respond to lessons planned using Aila compared to those [in the same subject] planned using other methods (GenAI and non-GenAI)?
- How do pupils respond to lessons and teaching materials planned using Aila?
- What is the perceived impact of using Aila on teachers' confidence [to plan and teach lessons]?
- To what extent, if at all, has using Aila influenced their broader planning and teaching practice? In what way(s) (for example, in structuring lessons or the activities used)?
- Do teachers intend to use (a) Aila or (b) GenAI for lesson planning in the future? Why/why not?
- Were there any financial costs or savings for teachers and schools when using Aila?

We explored teacher perceptions of the impact of using Aila on their broader planning and pedagogy.

We compared teachers' perceptions of their lesson planning workload¹² at baseline before randomisation and at the end of implementation (for both groups), and explored whether Aila teachers perceive any 'trade-off' in lesson planning between saving time and enhancing quality. This is important because teachers may be reinvesting saved time in higher-value planning activities or dedicating similar planning time while finding the process easier or more rewarding. This complements the analysis of self-reported lesson planning workload in the Impact Evaluation. Where teachers did perceive a reduction in their lesson planning workload, we will describe how teachers use the saved time.

We explored teacher perceptions of the process of teaching with Aila plans and resources, lesson quality, and pupil learning and engagement. At the end of the trial (endpoint survey, Appendix N) we asked teachers about their intentions for using Aila in future terms, as a comparison point for the longitudinal survey (Appendix O) which explored teachers' continuing use.

IPE RQ5: How does teachers' experience and confidence moderate Aila use and workload reduction? [moderators]

- How does teacher experience and confidence affect the proportion of lessons for which Aila is used?
- How does teacher experience and confidence moderate changes in teacher workload?

We adapted TPACK (Technological Pedagogical Content Knowledge) scales (for example, Schmid, Brianza and Petko, 2020) for the trial context to measure teachers' confidence in subject knowledge, pedagogical knowledge, and using technology. We collected data on years of teaching experience. We explored how these different factors moderated teachers' use of Aila, and any reduction in workload.

¹² Using questions about workload adapted from TALIS.

Research methods

The IPE research methods are summarised in Table 8 below.

Table 8: IPE methods overview

			IPE dimension/research area					Data analysis methods
			Usual practice	Adherence	Fidelity, teacher responsiveness	Perceived impact	Moderators	
			IPE RQ1	IPE RQ2	IPE RQ3	IPE RQ4	IPE RQ5	
Data collection approaches and sample	Diary	All teachers	X	X			X	Descriptive statistics
	Surveys	All teachers	X	X	X	X	X	Multi-level modelling (moderator analysis) Factor analysis (confidence measures)
	Aila usage data	Anonymous teacher-level data		X	X			Descriptive statistics
	Visit: observation and teacher interview	5 Aila teachers			X	X		Thematic analysis
	Online interview	8 teachers who used Aila	X		X	X		
	Online interview	8 teachers who did not use Aila	X	X				

Surveys

The baseline survey collected:

- expected time spent planning KS2 lessons in a week;
- approach to planning lessons (for example, individual, collaborative, whole school);
- current usage of Aila and/or other GenAI tools (for lesson planning);
- use of Oak National Academy resources, at any time prior to the trial, and immediately prior to the trial;
- teacher confidence in subject knowledge, pedagogy, and technology (scales adapted from TPACK) (Schmid, Brianza and Petko, 2020);
- teacher confidence in key aspects of lesson planning that Aila intends to scaffold, as a baseline for comparison; and
- key characteristic data on teachers such as years of experience, year group(s) taught, and current role.

The endpoint survey collected:

- teacher use and perceptions of other GenAI tools;
- teacher use of Aila, any use of Oak National Academy during the trial period, teacher evaluations of the Aila process, perceived impact and financial costs relating to Aila use, if any (only for the Aila group); and
- teacher confidence in key aspects of lesson planning that Aila scaffolds, to explore changes since baseline.

The longitudinal survey explored teachers' experiences in the academic term following the trial period (January to March 2026). The survey scope was adapted at the end of the trial (January 2026) in response to the data already collected. As originally planned, it covered key teacher-level outcomes, including teacher workload and continued use of Aila. In addition, it included questions on teachers' use of other GenAI tools to provide further context for teachers' decisions about when to use Aila.

The survey response rates (as a proportion of randomised teachers) were 100% for the baseline survey (this was a prerequisite for randomisation), 67% for the endpoint survey (69% of Aila-assigned teachers and 65% of non-Aila assigned teachers), and 48% for the longitudinal survey (51% of Aila-assigned teachers and 45% of non-Aila assigned teachers).

Diaries

As described in the Impact Evaluation Methods, the weekly diary collected self-reported information about the proportion of planned lessons for which teachers used Aila, for both Aila and non-Aila teachers. This was used to describe adherence for teachers in the Aila and non-Aila groups.

Aila usage data

We collected anonymous teacher-level data from Oak National Academy for each participating trial teacher that used Aila during the trial. To do this, we shared each teacher's email with the Academy, including a small number of alternative emails used to access Aila that were provided by teachers in the endpoint survey. Of the emails shared, 72% of Aila-assigned teachers and 14% of non-Aila assigned teachers were matched as being present in the Oak National Academy dataset. For Aila-assigned teachers, this is lower than the proportion who reported via diaries that they had used Aila (81%), suggesting that the shared usage data did not capture all usage from the trial. For example, if teachers from either group used multiple or alternative emails to access Aila but did not provide this email via the endpoint survey, their data would not be included in the dataset.

The shared data included the number of lessons started, the number of lessons downloaded, lesson subjects and topics, year group or key stage, and the number of refinements per lesson and time spent in each interaction. This enabled calculation of group averages as well as contextualisation of teacher case study and interview data.

Interviews and visits

In total, across visits and interviews, we collected qualitative data from 21 teachers across 20 schools, including a range of regions, school types, FSM-eligibility levels, and Ofsted ratings.

We conducted five visits with Aila teachers, comprising a lesson observation and teacher interview. While we recruited six schools for visits, as intended in the study design, one withdrew at short notice due to staff absence. Due to the short timeframe of this Teacher Choices trial it was not possible to schedule a replacement visit within the trial period.

Visited teachers were sampled purposively using baseline survey data and early diary data. As the visits focused on seeing Aila used in practice, and on the perceived impacts on teachers and pupils, we sampled purposively for teachers who had used Aila in at least two weeks across weeks one to three of the trial, and across a range of other reported GenAI use (for an average of zero to 50% of lessons planned across weeks one to three), and reported lesson planning hours (average three to ten hours per week across weeks one to three). As the visit sample is small and intentionally selected teachers with regular use of Aila in the first three weeks of the trial, it is not representative of the overall population of trial teachers.

In each school that was visited, we observed a lesson which was planned using Aila to explore pupil responsiveness based on observer notes. Before the lesson, we asked the teacher to identify any key differentiating features of the Aila lesson compared with non-Aila lessons for that subject. We focused on these features to collect field notes on teacher-pupil interactions and observed pupil responses to these features. We also photographed three to five anonymous samples of pupil work from each lesson. We interviewed the teacher after the lesson to understand their previous lesson planning practice, how and why they used Aila to plan the lesson, their experience of teaching the lesson, and their perception of pupil responsiveness and the impact of the lesson.

In addition, after the trial period we conducted online interviews with 16 teachers, excluding teachers from visited schools. This included eight who had used Aila during the ten weeks of the trial (seven from the Aila-assigned group, and one from the non-Aila assigned group) and eight who had not used Aila (all from the non-Aila assigned group).

Interviewees were sampled purposively using data from baseline and endpoint surveys and lesson planning diaries. In the Aila-assigned group, a key aim for interviews was to understand who Aila was working well for. Teachers were selected from those who regularly used Aila through the trial, reported a positive outcome (such as saving time), or intended to continue using Aila after the trial. The sample included teachers who were using other GenAI tools for lesson planning and teachers who were not. In the non-Aila assigned group, we sampled for a range of usual lesson planning approaches, including teachers who did not use GenAI, those who used education-specific GenAI tools (including one teacher who had used Aila), and those who used other GenAI tools.

Interviews with teachers who had used Aila focused on usual practice, how and why they had used it, and perceived impact. Interviews with teachers who had not used Aila focused on their usual practice in lesson planning, including their views and experience of using GenAI tools for lesson planning, and reasons for adherence or non-adherence.

Cost data

We anticipated that the main costs and savings related to Aila would be teachers' time (captured through the impact analysis), however, we checked the assumption of no direct financial costs by asking about any costs incurred in interviews and the endpoint survey.

Analysis

Qualitative data

Qualitative data from the interviews (n = 16) and visits (n = 5) were analysed thematically in MaxQDA, starting with deductive coding using the IPE dimensions as a coding frame and inductively coding or summarising text within each dimension. Using a deductive frame for top-level coding ensured that coding was focused on the research questions, while inductive coding within that frame ensured that all data relevant to the research questions was captured. The inductive coding was then compared with the logic model, including developing a context-specific description of the COM-B model (Michie, van Stralen and West, 2011), mapping teachers' perceptions of their confidence, opportunity, and motivation and their subsequent use of Aila. This comparison was completed after inductive coding to minimise selective coding based on the expected outputs and outcomes from the logic model. We established an audit trail by keeping copies of field notes, transcribed data, coded data, and case summaries, which were available to colleagues involved in quality assurance of the IPE.

Quantitative data—diaries, surveys and Aila usage data

We summarised the diary, survey, and Aila usage data with descriptive statistics, including frequencies, distribution statistics where appropriate (for example, mean, median, quartiles, min/max), and selected cross-tabulations. To monitor changes during the trial (in, for example, confidence), we used the matched sample of teachers who responded to both baseline and endpoint surveys. As the Aila usage data is anonymous, it was not possible to match this to survey and diary responses at teacher-level. Instead, we triangulated the overall results from the Aila usage data with diary and survey responses to understand how teachers' self-reported planning activity maps to the user data collected by Aila.

IPE RQ4.1: What is the impact of using Aila on teachers' self-reported perceptions of their workload?

In both baseline and endpoint surveys, teachers were asked, 'Was the amount of time you spent on lesson planning too little, too much or about right?' All respondents who answered both surveys were categorised as 'reduced', 'maintained', or 'increased' perceived workload depending on the difference between their baseline and endpoint responses. This variable was fit as an ordinal outcome in a cumulative link mixed model with intervention group as the key explanatory variable, controlling for baseline lesson planning time and stratification group by means of covariates and accounting for clustering of teachers within schools. The intervention group coefficient was tested for significance at the 5% level.

Quantitative data—survey moderator analysis

We undertook moderator analysis using the TPACK scale (Schmid, Brianza and Petko, 2020) and length of time teaching.

We looked at the relationship between confidence and length of time teaching, extent of use of Aila, and changes in workload. We first conducted a polychoric factor analysis on the items in the baseline teacher survey relating to length of time teaching and confidence. The relative loading of each variable onto the factors was interpreted alongside the crosstabs involving that variable (described in the Qualitative Data section above). After establishing the factor structure, we used the factors (details of the emerging factors can be found in the IPE Findings section) to answer the following questions:

(1) How is length of time teaching and confidence associated with the proportion of lessons for which Aila is used?

To answer this question, we built a binomial multilevel model with a logit link function to model how the proportion of lessons for which Aila is used was affected by baseline teacher factors in experience and confidence. The model accounted for the clustering of teachers in schools and the clustering of weeks within teachers by modelling them as random effects. The regression model was defined by

$$\text{logit}(pAila_{ijk}) = \sum_{f=1}^N \beta_f \text{factorscore}_{fjk} + \beta_{0k} + \beta_{0jk} + \beta_2 \text{BaseAiluse}_{jk}$$

where:

$pAila_{ijk}$ is the probability that teacher j in school k used Aila for planning in week i ; this was measured using a question from the weekly teacher diary and parameterised as a number of 'successes': (1) if Aila was used for lesson planning in a given week; and 'failures': (0) if Aila was not used in lesson planning in a given week for each teacher;

factorscore_{fjk} is the factor score for factor f for teacher j in school k ;

BaseAiluse_{jk} is a measure of how much teacher j in school k used generative AI prior to the study as measured by the teacher baseline survey;

β_{0k} is the intercept in school k (modelled as a random effect);

β_{0jk} is the intercept of teacher j in school k (modelled as a random effect);

β_f is the coefficient of interest estimating the association between factor f and the proportion of lessons planned using Aila;

N is the number of factors (or different dimensions) identified in the previous factor analysis; and

β_2 is a coefficient estimating the association between baseline teacher AI use and the proportion of lessons planned using Aila.

(2) How do length of time teaching and confidence moderate changes in teacher workload?

We built a series of linear multilevel models (one for each factor) similar to that used in the primary analysis but with the addition of main effect and interaction terms to measure moderating effect of each factor, controlling for baseline lesson planning time by means of a covariate and accounting for clustering of teachers at school level by modelling it as a random effect. Each teacher-level regression model was defined by

$$Workload_{ijk} = \beta_{0k} + \beta_{0jk} + \beta_1 Aila_k + \beta_2 BaseWorkload_{jk} + \beta_3 factorscore_{jk} + \beta_4 factorscore_{jk} * Aila_k + \epsilon_{ijk}$$

where:

$Workload_{ijk}$ is the lesson planning workload (in hours) of teacher j in school k in week i . This was measured using a weekly teacher diary;

$BaseWorkload_{jk}$ is the estimated lesson planning workload for teacher j in school k as measured by the baseline teacher survey;

β_{0k} is the intercept in school k (modelled as a random effect);

β_{0jk} is the intercept of teacher j in school k (modelled as a random effect);

β_1 is a coefficient of interest estimating how much difference being assigned to the Aila group makes to lesson planning workload;

β_2 is a coefficient estimating the association between baseline lesson planning workload and workload during the trial;

$Aila_k$ is a variable indicating whether school k was randomly assigned to use Aila or not;

$factorscore_{jk}$ is the factor score for teacher j in school k ;

β_3 is a coefficient of interest estimating the association between the confidence factor and lesson planning workload as a main effect;

β_4 is a coefficient of interest estimating how the interaction between the confidence factor score and group assignment affects lesson planning workload; and

ϵ_{ijk} is the residual error term for teacher j in school k in week i .

Synthesis of IPE data

We collated and triangulated data sources to provide a rounded picture of how both lesson planning approaches are enacted, and variation across schools and teachers. This included identifying any patterns based on key moderators. The IPE findings were used to contextualise impact findings and aid their interpretation. We used the data from the IPE, along with the impact findings, to update the theory of change (see Conclusions).

Timeline

All trial activities are undertaken by NFER unless otherwise specified below.

Table 9: Timeline of activities for the trial

Dates	Activity
January–May 2025	Project set-up meetings—NFER, EEF, Oak National Academy Finalise recruitment documents—NFER, EEF, Oak National Academy
June–July 2025	Promote the trial on website and social media—NFER and EEF School recruitment via EoI and MoUs Teacher data collection First Study Advisory Board meeting
September 2025	Teachers complete baseline survey Publish Study Plan Randomisation NFER informs participants about their group allocation and shares trial guidance documents
Late September–December 2025	Teachers adopt allocated ‘choice’ for ten weeks (between 27 September and 5 December 2025) Teachers complete weekly teacher diary Collection of lesson plans and resources IPE case study visits
December 2025–February 2026	Teachers complete end point survey Review of lesson resources (RQ2) Online teacher interviews Trial analysis commences
March 2026	Teachers complete longitudinal survey Emerging findings meeting (excluding the longitudinal survey) with the Study Advisory Board
April–May 2026	Longitudinal survey analysis Report writing First draft report submitted to the EEF
June–September 2026	Report revisions Report published on the EEF website
November 2026	NFER shares data with FFT for archiving

Impact evaluation results

Participant flow including losses and exclusions

The flow of participating teachers and schools through the trial can be seen in Figure 3 below.

NFER was responsible for school recruitment, which was completed between 15 May and 18 July 2025. We approached 15,755 primary schools in England with an EoI form. Of these, 248 schools that completed the EoI and met the eligibility criteria were subsequently contacted with a MoU. One additional school that completed the EoI was deemed ineligible because it wished to continue using the Oak National Academy website during the trial period and therefore was not asked to sign the MoU. In total, 141 schools completed the MoU and one withdrew before NFER approached schools for teacher data.

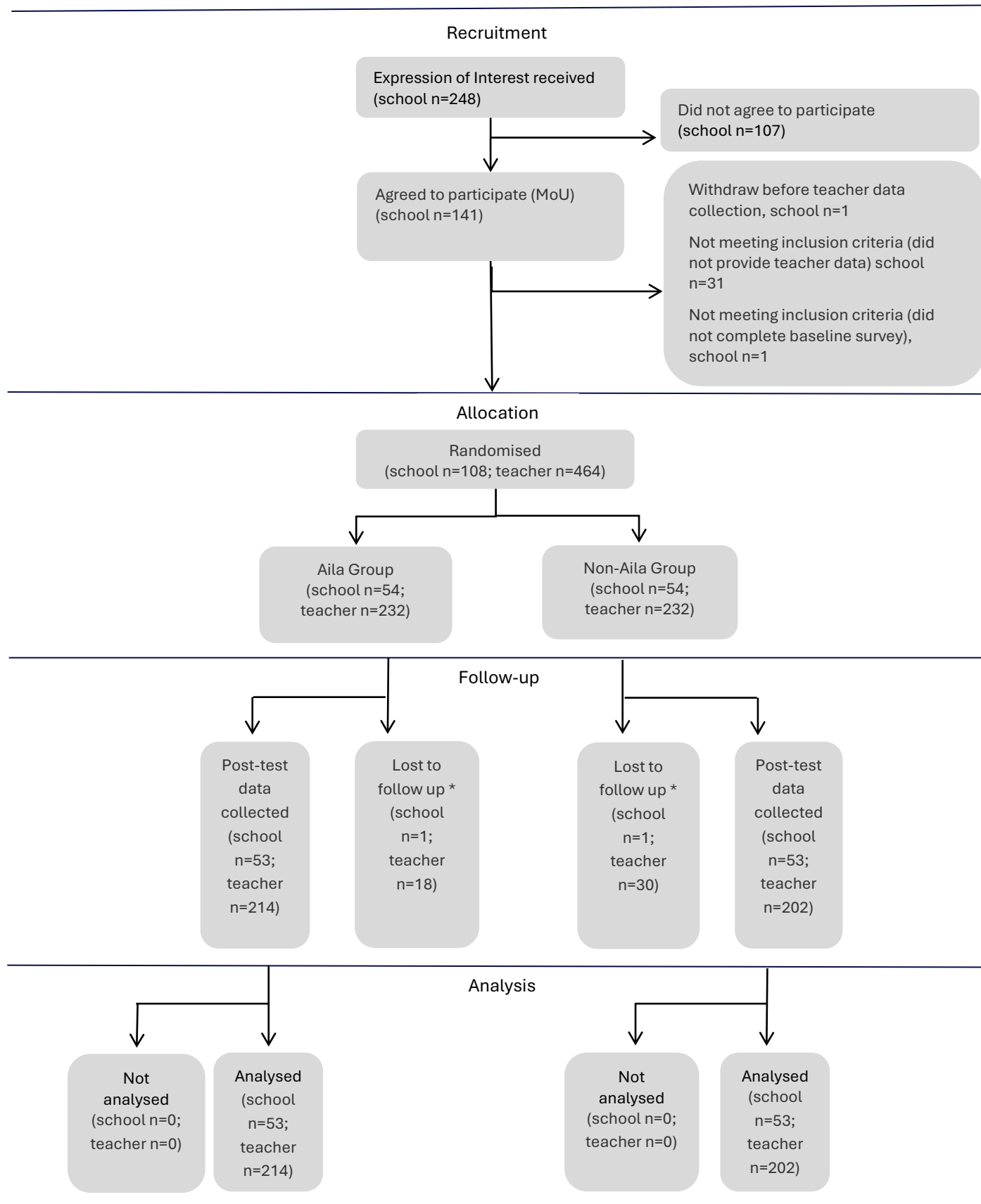
NFER contacted the key contact at each school to request the names and contact details of nominated teachers who would participate in the trial. Schools were sent up to two reminders to submit this data to NFER. In total, 109 schools provided teacher data before the end of the summer term.

At the end of summer 2025, 549 teachers from 109 schools were recruited to the trial. Prior to completion of the baseline survey, one school, involving two teachers, withdrew due to limited capacity to participate. Of the 547 teachers invited to complete the baseline survey, 464 completed it and were subsequently randomised. The remaining 83 teachers who did not complete the baseline survey were informed that they would not be progressing in the trial.

All 108 schools that provided their teacher data and had at least one nominated teacher who completed the survey by the deadline became part of the sample that was randomised on 23 September 2025. Schools were randomised at school level, stratified by the number of teachers with no/low prior GenAI use, into either the Aila group or the non-Aila group: 232 teachers were randomised to the Aila group and 232 to the non-Aila group.

Of the 464 participating teachers at randomisation, 177 were also the trial lead. Trial Leads and participating teachers were informed of their randomisation results of the 23 September 2025. Within this email, teachers were provided with trial guidance documents as per the random allocation.

Figure 3: Participant flow diagram



* For lost-to-follow-up numbers, please note that the school n is the number of schools completely missing and the teacher n is also across many other schools.

Table 10 shows the minimum detectable effects sizes described in the Sample Size section above. The MDES at analysis stage was smaller than the MDESs at design and randomisation. This is driven by smaller than anticipated ICCs at both teacher and school level, while the total number of teachers at analysis was more than expected at design or randomisation (numbers in brackets in the table) and the average number of weeks was also higher at analysis. Conversely, the proportion of variance explained by teacher- and school-level covariates was not as high as anticipated, but the detrimental effect of this was not evident in the MDES. As described in the Methods, the minimum detectable fold change (MDFC) is estimated as $10^{(\text{MDES} \times \text{Standard Deviation})}$. The standard deviation in the ChatGPT trial (on the $\log_{10}$ scale) was 0.424 and in the current trial was 0.550, so while the MDES was lower at the analysis stage, the MDFC was slightly higher.

Table 10: Minimum detectable effect size at different stages

Values in brackets indicate MDES and numbers after accounting for assumed attrition between randomisation and analysis.

		Design	Randomisation	Analysis
Minimum Detectable Effect Size (MDES)		0.220 (0.239)	0.223 (0.241)	0.194
Minimum Detectable Fold Change		1.24 (1.26)	1.24 (1.27)	1.28
Proportion of variance explained by covariates (R²)	Level 2 (teacher)	0.06	0.06	0.04
	Level 3 (school)	0.40	0.40	0.16
Intraclass correlations (ICCs)	Level 2 (teacher)	0.40	0.40	0.27
	Level 3 (school)	0.11	0.11	0.05
Alpha		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided
Average cluster size (teachers)		3.2 (2.7)	4.3 (3.7)	3.9
Average cluster size (weeks)*		8.2 (8.2)	8.2 (8.2)	9.0
Number of schools	Aila Group	67 (63)	54 (52)	53
	Non-Aila Group	66 (63)	54 (51)	53
	Total:	133 (126)	108 (103)	106
Number of teachers	Aila Group	213 (172)	232 (188)	214
	Non-Aila Group	213 (172)	232 (187)	202
	Total:	426 (344)	464 (375)	416

* The assumed number of weeks of diary completed by an average teacher.

Attrition

Table 11: Teacher-level attrition from the trial (primary outcome)

		Aila	Non-Aila	Total
Number of teachers	Randomised	232	232	464
	Analysed	214	202	416
Teacher attrition (from randomisation to analysis)	Number	18	30	48
	Percentage	7.8%	12.9%	10.3%

Of the 464 teachers who were randomised, 48 (or 10.3%) were randomised but did not complete any diary entry. Attrition was slightly higher in the Non-Aila group, but this difference was not significant (see Missing Data section). We requested feedback on the reasons for attrition, and the responses were ‘capacity issues’ and ‘teacher has left the school’.

Teacher and school characteristics

Table 12 details school and teacher characteristics for the two study groups at randomisation. Table 13 shows the same characteristics for the schools and teachers that were included in the primary analysis (analysed groups). For continuous measures, the Cohen’s *d* effect size is reported (for Aila minus Non-Aila) along with the *p*-value for an independent sample *t*-test. For school-level characteristics, the national mean is calculated from a ‘get information about schools’¹³ dataset downloaded on the date of randomisation and subset to only include schools where the PhaseOfEducation (variable name) was one of ‘All-through’, ‘Middle deemed primary’, or ‘Primary’ and the EstablishmentStatus (variable name) was ‘Open’ or ‘Open, but proposed to close’. This resulted in *N* = 16,896. For two of the teacher-level characteristics, the national mean is calculated from the TALIS workload survey (OECD, 2018) using the primary school teachers in England subset.

Overall, at randomisation the sample of schools appears well balanced between the groups. The most notable difference between arms appears to be in regional distribution (GOR) where more of the Aila group are located in the South West and more of the Non-Aila group are located in the East Midlands and West Midlands.

The sample of teachers also appears generally well balanced, although the Aila group appears to have more senior leaders than the Non-Aila group.

None of the *p*-values from the independent sample *t*-tests were less than 0.05.

Table 12: Baseline characteristics of groups as randomised

School level (categorical)	National level %	Aila Group		Non-Aila Group	
		n/N (missing)	%	n/N (missing)	%
GOR: East Midlands	9.90%	3/54 (0)	5.6%	9/54 (0)	16.7%
GOR: East of England	11.80%	6/54 (0)	11.1%	2/54 (0)	3.7%
GOR: London	10.70%	5/54 (0)	9.3%	8/54 (0)	14.8%
GOR: North East	5.10%	5/54 (0)	9.3%	2/54 (0)	3.7%
GOR: North West	14.50%	6/54 (0)	11.1%	2/54 (0)	3.7%

¹³ <https://www.get-information-schools.service.gov.uk/>

GOR: South East	15.50%	13/54 (0)	24.1%	4/54 (0)	7.4%	
GOR: South West	11.20%	4/54 (0)	7.4%	7/54 (0)	13.0%	
GOR: West Midlands	10.60%	8/54 (0)	14.8%	13/54 (0)	24.1%	
GOR: Yorkshire and the Humber	10.50%	4/54 (0)	7.4%	7/54 (0)	13.0%	
Urban/Rural: Larger rural	10.40%	5/54 (0)	9.3%	7/54 (0)	13.0%	
Urban/Rural: Smaller rural	17.60%	9/54 (0)	16.7%	5/54 (0)	9.3%	
Urban/Rural: Urban	72%	40/54 (0)	74.1%	42/54 (0)	77.8%	
Ofsted Rating: * Outstanding		10/54 (0)	18.5%	10/54 (0)	18.5%	
Ofsted Rating: Good		38/54 (0)	70.4%	40/54 (0)	74.1%	
Ofsted Rating: Requires improvement or Not judged **		6/54 (0)	11.1%	4/54 (0)	7.4%	
School level (continuous)	National-level mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size; p -value
Percentage FSM	24.5 (15.8)	53 (1)	23.8 (16.1)	54 (0)	22.1 (14.6)	ES = 0.11; p = 0.567
KS2 Reading Average	105.6 (2.8)	52 (2)	106 (2.6)	53 (1)	106.2 (2.7)	ES = -0.07; p = 0.715
KS2 Maths Average	104.6 (3)	52 (2)	105.1 (3.2)	53 (1)	105.1 (2.9)	ES = 0.01; p = 0.973
Teacher level (categorical)		n/N (missing)	%	n/N (missing)	%	
Role: Classroom teacher		122/231 (1)	52.8%	129/232 (0)	55.6%	
Role: Middle leader (e.g. subject or phase leader)		69/231 (1)	29.9%	72/232 (0)	31.0%	
Role: Senior leader (including headteacher)		33/231 (1)	14.3%	22/232 (0)	9.5%	
Role: Other		7/231 (1)	3.0%	9/232 (0)	3.9%	
Prior GenAI Use: Most/all lessons		15/232 (0)	6.50%	12/232 (0)	5.20%	
Prior GenAI Use: At least one lesson a week		72/232 (0)	31%	77/232 (0)	33.20%	
Prior GenAI Use: Less often		89/232 (0)	38.40%	68/232 (0)	29.30%	
Prior GenAI Use: I have never/I no longer use		56/232 (0)	24.10%	75/232 (0)	32.30%	

GenAI to support my lesson planning						
Trial Lead: No		201/232 (0)	86.60%	208/232 (0)	89.70%	
Trial Lead: Yes		31/232 (0)	13.40%	24/232 (0)	10.30%	
Teacher level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size; p-value
Year(s) working as a teacher (total)	12.1 (8.9)	231 (1)	11.7 (7.6)	232 (0)	10.9 (7.6)	ES = 0.1; p = 0.273
Year(s) working as a teacher at current school	6.9 (6.5)	231 (1)	7.1 (6)	231 (1)	6.5 (5.9)	ES = 0.09; p = 0.31
Baseline question from baseline survey: 'Approximately how many hours per week do you expect to spend on lesson planning for KS2 classes this term?'		232 (0)	9.4 (9)	232 (0)	8.8 (6.9)	ES = 0.07; p = 0.484
Baseline question from baseline survey: 'Approximately how many hours per week do you expect to spend on lesson planning for KS2 classes this term?' (log ₁₀ transformed)		232 (0)	0.8 (0.3)	232 (0)	0.8 (0.3)	ES = -0.03; p = 0.775

* Ofsted rating is no longer published in the GIAS (<https://www.get-information-schools.service.gov.uk/>) dataset so we have no national average for these categories.

** Combined due to low number of schools in at least one of these categories.

Table 13: Baseline characteristics of groups at analysis

School level (categorical)	National level %	Aila Group		Non-Aila Group	
		n/N (missing)	%	n/N (missing)	%
GOR: East Midlands	9.90%	3/54 (0)	5.6%	9/54 (0)	16.7%
GOR: East of England	11.80%	6/54 (0)	11.1%	2/54 (0)	3.7%
GOR: London	10.70%	5/54 (0)	9.3%	8/54 (0)	14.8%
GOR: North East	5.10%	5/54 (0)	9.3%	2/54 (0)	3.7%
GOR: North West	14.50%	6/54 (0)	11.1%	2/54 (0)	3.7%
GOR: South East	15.50%	13/54 (0)	24.1%	4/54 (0)	7.4%
GOR: South West	11.20%	4/54 (0)	7.4%	7/54 (0)	13.0%

GOR: West Midlands	10.60%	8/54 (0)	14.8%	13/54 (0)	24.1%	
GOR: Yorkshire and the Humber	10.50%	4/54 (0)	7.4%	7/54 (0)	13.0%	
Urban/Rural: Larger rural	10.40%	5/54 (0)	9.3%	7/54 (0)	13.0%	
Urban/Rural: Smaller rural	17.60%	9/54 (0)	16.7%	5/54 (0)	9.3%	
Urban/Rural: Urban	72%	40/54 (0)	74.1%	42/54 (0)	77.8%	
Ofsted Rating: Outstanding		10/54 (0)	18.5%	10/54 (0)	18.5%	
Ofsted Rating: Good		38/54 (0)	70.4%	40/54 (0)	74.1%	
Ofsted Rating: Requires improvement or Not judged		6/54 (0)	11.1%	4/54 (0)	7.4%	
School level (continuous)	National-level mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size; p-value
Percentage FSM	24.5 (15.8)	53 (1)	23.8 (16.1)	54 (0)	22.1 (14.6)	ES = 0.11; p = 0.567
KS2 Reading Average	105.6 (2.8)	52 (2)	106 (2.6)	53 (1)	106.2 (2.7)	ES = -0.07; p = 0.715
KS2 Maths Average	104.6 (3)	52 (2)	105.1 (3.2)	53 (1)	105.1 (2.9)	ES = 0.01; p = 0.973
Teacher-level (categorical)		n/N (missing)	%	n/N (missing)	%	
Role: Classroom teacher		110/213 (1)	51.6%	111/202 (0)	55.0%	
Role: Middle leader (e.g. subject or phase leader)		65/213 (1)	30.5%	66/202 (0)	32.7%	
Role: Senior leader (including headteacher)		32/213 (1)	15.0%	17/202 (0)	8.4%	
Role: Other		6/213 (1)	2.8%	8/202 (0)	4.0%	
Prior GenAI Use: Most/all lessons		15/214 (0)	7%	9/202 (0)	4.50%	
Prior GenAI Use: At least one lesson a week		67/214 (0)	31.30%	69/202 (0)	34.20%	
Prior GenAI Use: Less often		81/214 (0)	37.90%	57/202 (0)	28.20%	
Prior GenAI Use: I have never/I no longer use GenAI to support my lesson planning		51/214 (0)	23.80%	67/202 (0)	33.20%	
Trial Lead: No		184/214 (0)	86%	179/202 (0)	88.60%	

Trial Lead: Yes		30/214 (0)	14%	23/202 (0)	11.40%	
Teacher-level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size; <i>p</i> -value
Year(s) working as a teacher (total)	12.1 (8.9)	213 (1)	11.8 (7.6)	202 (0)	10.8 (7.7)	ES = 0.13; <i>p</i> = 0.202
Year(s) working as a teacher at current school	6.9 (6.5)	213 (1)	7.2 (6.1)	201 (1)	6.6 (6.2)	ES = 0.1; <i>p</i> = 0.315
Baseline question from baseline survey: 'Approximately how many hours per week do you expect to spend on lesson planning for KS2 classes this term?'		214 (0)	9.2 (8.9)	202 (0)	9.1 (7)	ES = 0; <i>p</i> = 0.972
Baseline question from baseline survey: 'Approximately how many hours per week do you expect to spend on lesson planning for KS2 classes this term?' (log ₁₀ transformed)		214 (0)	0.8 (0.3)	202 (0)	0.9 (0.3)	ES = -0.1; <i>p</i> = 0.317

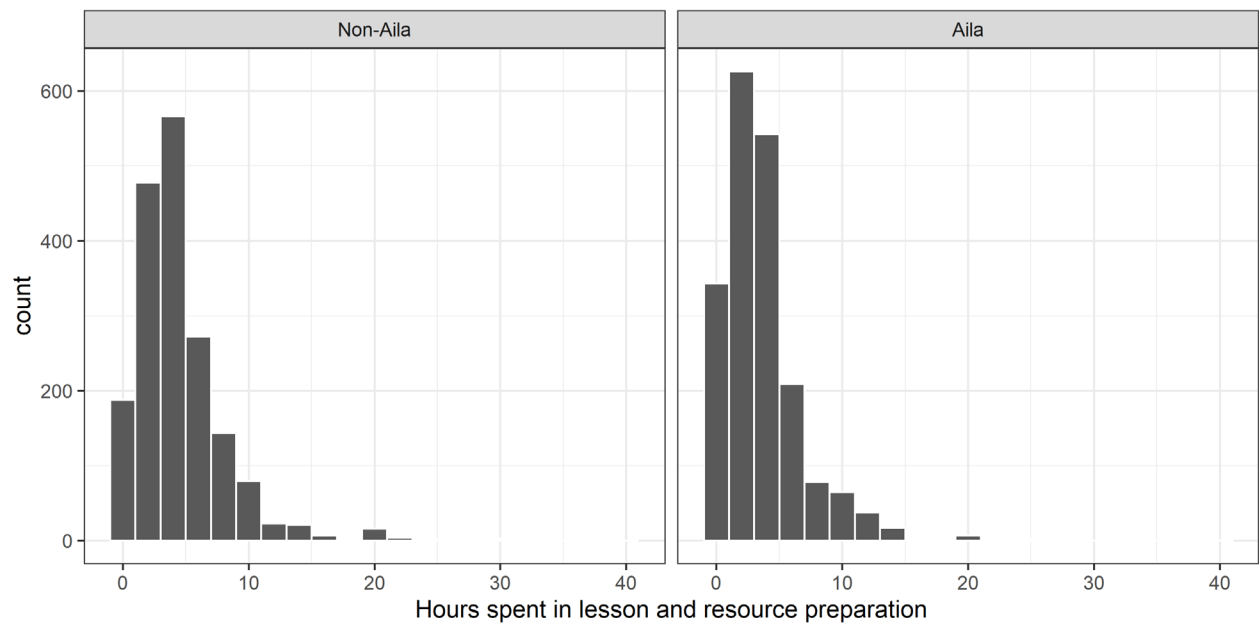
Outcomes and analysis

Primary analysis

RQ1: What is the impact of Aila on the average time spent per week on lesson planning for KS2 lessons during the ten-week trial period?

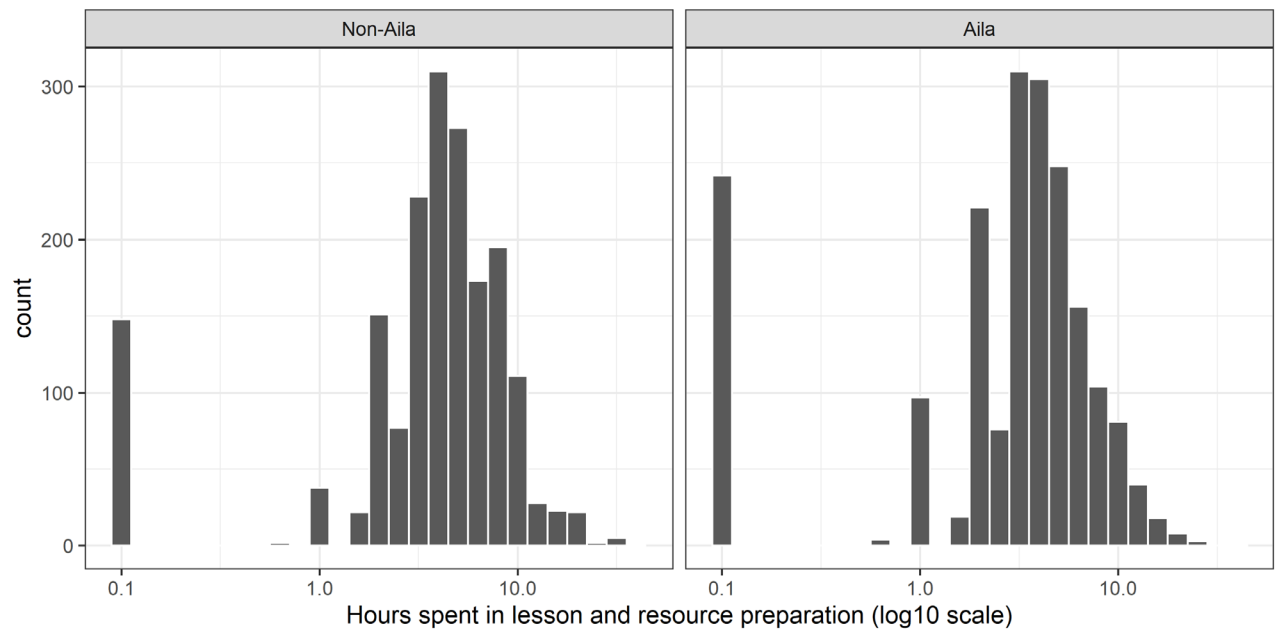
As specified in the study plan, we fitted the primary analysis model and visually inspected the residual plots for variance heterogeneity. This inspection revealed heteroscedasticity. Figure 4 presents the primary outcome distribution (weekly hours spent in lesson and resource preparation), which demonstrates the skewedness of the data.

Figure 4: Histogram of raw primary outcome data (weekly hours spent in lesson and resource preparation)



Therefore, as anticipated in the study plan, we applied a $\log_{10}$ transformation to the outcome measure (with a +0.1 offset to allow inclusion of zeros) before refitting the model. In the transformed model, although the zeros still caused some non-normality, overall the residual distribution was much better. It is documented (Schmidt and Finan, 2018) that only strong deviations from a Gaussian distribution of residuals increase the risk for biased conclusions so after transformation, the assumption of constant variance was considered upheld. The transformed outcome measure was also used in all secondary analyses. The distribution of the weekly hours spent in lesson and resource preparation after the $\log_{10}$ transformation is shown in Figure 5.

Figure 5: Histogram of $\log_{10}$ transformed primary outcome data (weekly hours spent in lesson and resource preparation)



The distribution of the baseline measure for RQ1 followed a similar distribution to the outcome measure (histograms in Appendix B). This variable was also $\log_{10}$ transformed to maximise the variance explained by the baseline in the analysis model.

Table 14 presents unadjusted means¹⁴ and confidence intervals for the primary outcome (calculated on the log₁₀ scale before back transformation), along with the fold change and effect size calculated from the model. The fold change indicates that the time spent on lesson planning in the Aila group was 76% of the time spent on planning in the Non-Aila group with a 95% confidence interval of (64%, 89%). Using the adjusted means from the model (Table 23), a typical teacher in the Aila group spent 2.51 hours a week (2.21, 2.86) preparing lessons and resources, while a typical teacher in the Non-Aila group spent 3.32 hours a week (2.90, 3.79), representing a saving of 0.81 hours (48.6 minutes) per week.

Table 14: Primary analysis (RQ1)

	Unadjusted means				Effect size			
	Aila group		Non-Aila group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Weekly hours spent in lesson and resource preparation	214 (18)	2.46 (2.32, 2.61)	202 (30)	3.33 (3.15, 3.51)	416 (48)	-0.22 (-0.35, -0.09)	0.76 (0.64, 0.89)	0.001

Secondary analysis

RQ2: How does the quality of lesson plans and resources in the Aila group compare to the quality of lesson plans and resources in the non-Aila group?

Reliability metrics

Interrater reliability, calculated as percentage agreement, was low: 8.3% overall, 6.7% in the Non-Aila group and 10% in the Aila group. This calculation represents the IRR metric specified in the study plan but this quite strict metric does not feel appropriate, with hindsight, because the outcome score is continuous (scale: 0 to 25) so perfect agreement between reviewers is not to be expected. As described in the methods, we explored the following alternative options:

- use all sampled lesson resources, not just those in the ‘reliability pool’;
- allow a tolerance for agreement, that is, the difference between scorers must be ≤ 2 ; and
- calculate the intra-class correlation coefficient (ICC).

Table 15 details the agreement scores for each of the options.¹⁵ Please recall that these are all post-hoc specified calculations.

¹⁴ Unadjusted means are calculated from the raw data whereas adjusted means are taken from model output and calculated for a teacher with average covariate values at an average school. Unadjusted means may be more prone to unseen biases in the school sample analysed (for example, more influence given to school practices in a school with many teachers in this study) so for preference, adjusted means should be interpreted.

¹⁵ Cicchetti (1994) gives the following guidelines for interpretation for kappa or ICC inter-rater agreement measures: less than 0.40—poor, between 0.40 and 0.59—fair, between 0.60 and 0.74—good, and between 0.75 and 1.00—excellent.

Table 15: Reliability metrics

Reliability score method	Overall	Aila Group	Non-Aila Group
Only reliability pool, perfect agreement (as per study plan)	0.08	0.10	0.07
Using all resources, perfect agreement	0.08	0.10	0.06
Using all resources, agreement with a tolerance ± 2	0.50	0.48	0.51
Using all resources, ICC	0.50	0.26	0.70

The percentage agreement is ‘poor’ when we use perfect agreement and ‘fair’ when we allow a ± 2 tolerance. The ICC is ‘fair’ overall but ‘poor’ in the Aila group and ‘good’ in the Non-Aila group. The ICC takes into account the between lesson plan variability relative to the total of the within lesson plan and between lesson plan variability and a high ICC means that the majority of the variance is between lesson plans rather than within them (that is, between scorers in the same lesson plan). So the lower ICC in the Aila group could either indicate that reviewers are more likely to disagree on lesson quality for Aila group lessons (larger within lesson plan variability) or that the underlying quality of the Aila group lesson plans is similar within the reviewed set (smaller between lesson plan variability), or a combination of both. We also calculated the percentage (perfect) agreement within each of the five scoring criteria (each with a scale of 0 to 5). These are shown in Table 16.

Table 16: Proportion agreement between reviewers for each scoring criteria

Scoring Criterion	Overall	Aila Group	Non-Aila Group
Clarity and structure of lesson resources	0.41	0.38	0.44
Extent that activities involve pupils with their learning progression	0.50	0.46	0.53
Extent that activities check the understanding of the learners	0.35	0.37	0.33
Appropriateness for the age group and ability level of the class	0.46	0.43	0.49
Quality and accuracy of subject content	0.41	0.45	0.38

Agreement is naturally higher within these categories than for the overall score due to the shorter scale, but none of them stand out as particularly strong or weak in agreement. Overall, we would consider the agreement among reviewers to be fair. Variability among reviewers is to be expected and is accounted for in the analysis model. We have no evidence of bias among reviewers.

Branding bias check

The outcome of the branding bias check model is shown in Table 17. As can be seen from the table, the scores for the ‘soft’ and ‘strict’ de-branded lessons are not meaningfully different. The likelihood ratio test p -value for the inclusion of the interaction term was not significant, $p = 0.89$, which supports the same conclusion.

Table 17: Branding bias check model outcomes

		Adjusted means				Effect size			
		Aila group		Non-Aila group					
Outcome	Branding	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Difference (95% CI)	p-value
Lesson quality score	Soft	4 (0)	19.0 (16.7, 21.3)	4 (0)	15.8 (13.5, 18.0)	8 (0)	0.92 (0.09, 1.75)	3.25 (0.33, 6.17)	0.025
Lesson quality score	Strict	4 (0)	19.1 (16.8, 21.4)	4 (0)	15.6 (13.3, 17.9)	8 (0)	0.99 (0.16, 1.82)	3.50 (0.58, 6.42)	0.015

Analysis model

The outcome of the linear mixed model comparing scores between the two randomisation groups is shown in Table 18. The non-inferiority *p*-value is significant, indicating that the quality of the resources in the Aila group are not meaningfully worse than the quality of the resources in the Non-Aila group, using a limit of -2.5 (as described in the Methods section, the limit was increased from -2 to -2.5 to mirror the change from a scale of 0–20 to 0–25). This model also accounts for the variability brought by individual reviewers' scoring and therefore we can say that the quality of the lesson resources in the Aila group was not inferior to that of the Non-Aila group.

Table 18: Secondary analysis (RQ2)

	Unadjusted means				Effect size			
	Aila group		Non-Aila group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Difference (95% CI)	p-value
Lesson quality score	21 (0)	17.7 (16.9, 18.5)	21 (0)	16.7 (15.6, 17.8)	42 (0)	0.25 (-0.08, 0.58)	1.01 (-0.34, 2.37)	<0.001

Dilution check

As described in the methods, since more than 25% (29%) of the lesson resources in the Aila group were not created using Aila, two further analyses were undertaken. The non-Aila group resources were weighted using the following lesson and teacher characteristics. Teacher characteristics: number of years teaching experience, workload from the baseline survey, workload as a source of stress from baseline survey; lesson characteristics: target year group and subject. Target class ability was not used as a weighting variable as more than 90% of selected resources were noted as 'mixed ability class'. Propensity score weighting was applied at a lesson level to the Non-Aila group lessons and the RQ2 model was rerun with these weights applied and with only the Aila created Aila group lessons included. The outcome of this analysis is shown in Table 19. Similar to the non-weighted RQ2 analysis, the non-inferiority *p*-value is significant, indicating that the quality of the resources in the Aila group are not significantly worse than the quality of the resources in the Non-Aila group, using a limit of -2.5. The magnitude and direction of the effect are also similar to those seen in the main RQ2 analysis model.

Table 19: Weighted RQ2 analysis

	Unadjusted means				Effect size			
	Aila group		Non-Aila group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Difference (95% CI)	p-value
Lesson quality score	15 (0)	18.5 (17.4, 19.6)	21 (0)	16.8 (15.9, 17.7)	36 (0)	0.49 (0.13, 0.85)	1.70 (0.45, 2.95)	<0.001

Table 20 shows the outcomes of both stages of the two-stage, least squares instrumental variable model. The coefficient of the second stage was not statistically significant, which means we detected no difference in Total Score between compliers (those who created or would have created their lessons using Aila) and non-compliers.

Table 20: Instrumental variable analysis

Model Stage	Total number of reviews (Aila; Non- Aila)	Total number of teachers/lessons (Aila; Non-Aila)	Predictor	Coefficient (95% CI)	p-value
Stage 1: Created using Aila variable regressed on randomisation group	138 (69; 69)	42 (21; 21)	Randomisation Group	0.696 (0.586, 0.805)	<0.001
Stage 2: Total score regressed on predicted 'Created using Aila' variable	138 (69; 69)	42 (21; 21)	Predicted 'Created using Aila' variable	1.46 (-0.437, 3.35)	0.134

Together, these two pieces of additional analyses give us confidence that the equivalence declared between the quality scores for the Aila group and Non-Aila group lesson resources is not driven by low use of Aila in the Aila group.

Sensitivity analyses

RQ3: What is the impact of Aila on average time spent per week on lesson planning for KS2 lessons excluding the first week of Aila use? [to allow for familiarisation]

Table 21 presents unadjusted means and confidence intervals for the RQ3 analysis along with the fold change and effect size calculated from the model. The fold change indicates that the time spent on lesson planning in the Aila group was 72% of the time spent on planning in the Non-Aila group with a 95% confidence interval (61%, 85%). Using the adjusted means from the model (Table 23), when the first week of Aila use is excluded, a typical teacher in the Aila group spent 2.39 hours a week (2.09, 2.74) preparing lessons and resources, while a typical teacher in the Non-Aila group spent 3.32 hours a week (2.89, 3.81), representing a saving of 0.93 hours per week (55.8 minutes). This slightly greater effect than in RQ1 is driven by lower average lesson planning hours in the Aila group when excluding the first week of Aila use, suggesting higher average lesson planning hours in the first week of Aila use, in other words, a learning effect.

Table 21: Sensitivity analysis (RQ3)

	Unadjusted means				Effect size			
	Aila group		Non-Aila group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Weekly hours spent in lesson and resource preparation	211 (21)	2.35 (2.20, 2.51)	202 (30)	3.33 (3.15, 3.51)	413 (51)	-0.25 (-0.38, -0.12)	0.72 (0.61, 0.85)	<0.001

RQ4: What is the impact of Aila on average time spent per week on lesson planning for KS2 lessons excluding half term week(s)?
[to allow for different amounts of, and variability in, planning during holiday time]

Table 22 presents unadjusted means and confidence intervals for the RQ4 analysis along with the fold change and effect size calculated from the model. The fold change indicates that, when half term was excluded, the time spent on lesson planning in the Aila group was 72% of the time spent on planning in the Non-Aila group with a 95% confidence interval of (62%, 85%). Using the adjusted means from the model (Table 23), a typical teacher in the Aila group spent 2.88 hours a week (2.54, 3.27) preparing lessons and resources, while a typical teacher in the Non-Aila group spent 3.98 hours a week (3.49, 4.54), representing a saving of 1.1 hours per week (66 minutes). Both groups had higher average weekly lesson planning hours when excluding half term week(s) suggesting that, for an average teacher, less planning takes place during half term.

Table 22: Sensitivity analysis (RQ4)

	Unadjusted means				Effect size			
	Aila group		Non-Aila group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Weekly hours spent in lesson and resource preparation	214 (18)	2.8 (2.65, 2.96)	202 (30)	3.96 (3.79, 4.14)	416 (48)	-0.30 (-0.45, -0.15)	0.72 (0.62, 0.85)	<0.001

Summary findings—adjusted means

Table 23 contains the adjusted means for the outcomes in RQ1, RQ2, RQ3, and RQ4 for both randomisation groups. The means are extracted from the analysis models so are adjusted for the average baseline (log₁₀ anticipated hours of workload) for the average stratification group and for the clustering at school and teacher level.

Table 23: Adjusted means for RQ1 to RQ4 analyses

Research question	Outcome measure	Aila Group		Non-Aila Group	
		N	Adjusted Mean (95% CI)	N	Adjusted Mean (95% CI)
RQ1	Weekly hours spent in lesson and resource preparation	214	2.51 (2.21, 2.86)	202	3.32 (2.9, 3.79)
RQ2	Lesson quality score	21	17.7 (16.6, 18.8)	21	16.7 (15.6, 17.8)
RQ3	Weekly hours spent in lesson and resource preparation	211	2.39 (2.09, 2.74)	202	3.32 (2.89, 3.81)
RQ4	Weekly hours spent in lesson and resource preparation	214	2.88 (2.54, 3.27)	202	3.98 (3.49, 4.54)

Analysis in the presence of non-compliance

RQ6: What is the impact of using Aila on the average time spent per week on lesson planning for KS2 lessons among ‘compliant’ teachers during the ten-week trial period?

In this trial ‘compliance’ is considered a continuous variable, defined as the percentage of lessons in which Aila was used for lesson planning, averaged for each teacher. The distribution of this variable in each of the randomisation groups is shown in Figure 6.

Figure 6: Percentage of lessons planned using Aila

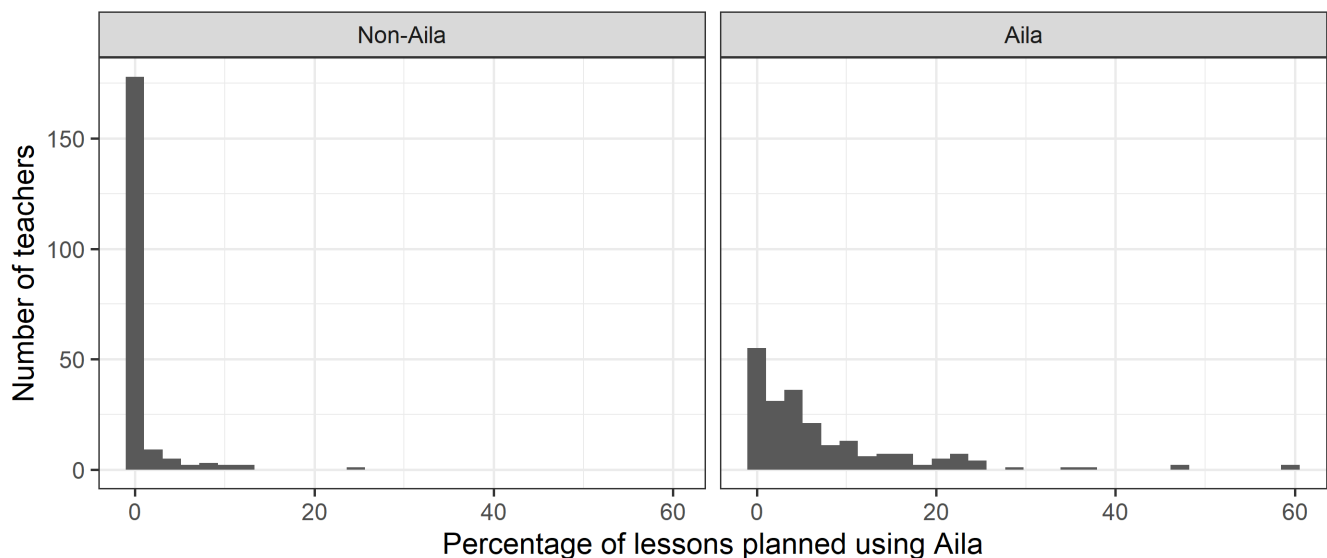


Table 24 shows the outcomes of both stages of the two-stage, least squares compliance model. For an increase of one unit (%) in predicted compliance we see a statistically significant 2% reduction in workload. Expanding this out to a larger change in compliance, for an increase of 10% in predicted compliance (i.e. using Aila to plan an additional 10% of lessons) the fold change becomes 0.83—a 17% reduction in workload.¹⁶

¹⁶ Calculated as $100 \times (1 - \text{'One Unit Fold Change'}^{10})$. Slight discrepancy due to rounding as one unit fold change = 0.98101.

Table 24: Compliance analysis

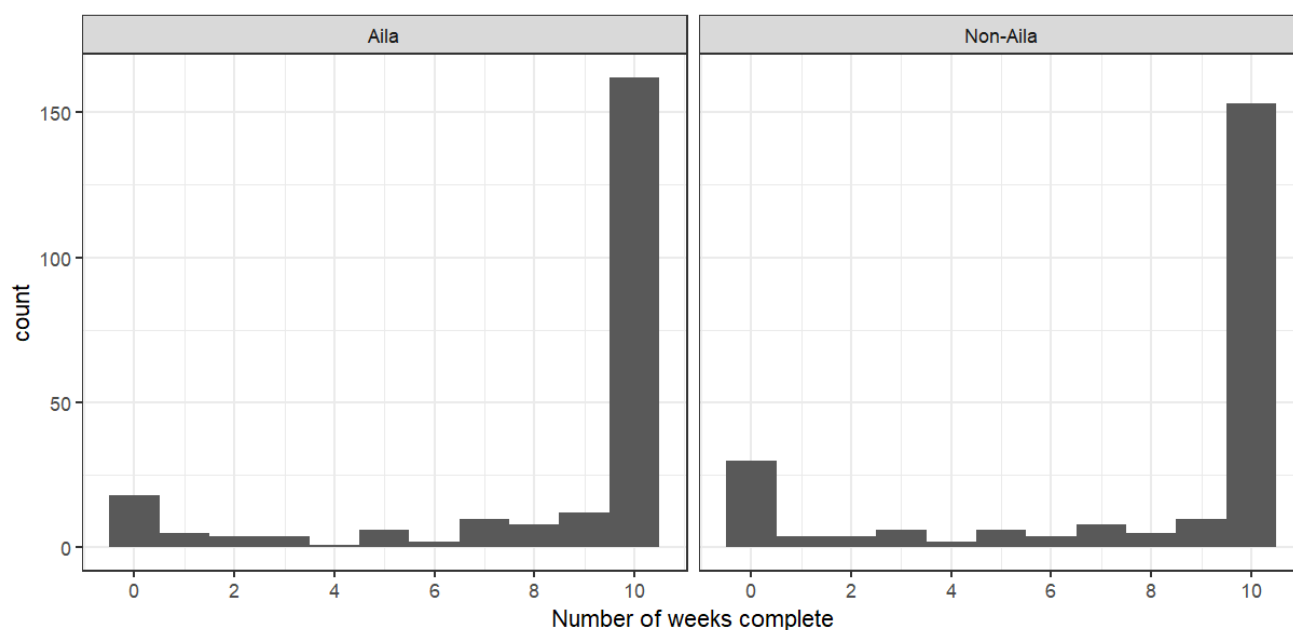
Model Stage	Total n (Aila; Non-Aila)	Predictor	Coefficient (95% CI)	p-value
Stage 1: Compliance variable regressed on randomisation group	416 (214; 202)	Randomisation Group	6.83 (5.46, 8.20)	<0.001
Stage 2: Workload (log10) regressed on compliance variable	416 (214; 202)	Compliance variable	0.98 (0.97, 0.99)	0.004

Missing data analysis

Table 25: Number of teachers with weekly diary entries

N weeks complete	N Aila Group Teachers	N Non-Aila Group Teachers	Total N Teachers
0	18	30	48
1	5	4	9
2	4	4	8
3	4	6	10
4	1	2	3
5	6	6	12
6	2	4	6
7	10	8	18
8	8	5	13
9	12	10	22
10	162	153	315

Figure 7: Number of teachers with weekly diary entries



Of the 464 teachers who were randomised, 48 (or 10%) did not complete any diary entry. These teachers were not included in any analysis but all other teachers were. The average number of weeks complete was eight. There was no evidence that ‘having any missing data’—one or more missing diary entries—differed by randomisation allocation, with the chi squared test p -value = 0.426.

When we fit the binomial generalised linear mixed effects model to investigate patterns of missingness, only ‘whether [participant is] trial lead’ was significantly associated with missingness. This association was negative, that is, if a teacher was the ‘trial lead’ for their school—the main contact for the research—they were less likely to be missing from the analysis.

Table 26 shows the results from the sensitivity analysis where ‘trial lead’ was included as a covariate in the primary model. The complete case analysis where the primary analysis was rerun using only the 315 teachers who completed all ten diaries, along with the primary analysis model results, are shown again for comparison. Results from both of these additional analyses were very similar to the primary analysis, which means no concerns of bias were raised by the missing data analysis.

Table 26: Outcomes from missing data sensitivity analysis and complete case analysis

		Unadjusted means				Effect size			
		Aila group		Non-Aila group					
Analysis	Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Sensitivity analysis with trial lead as covariate	Weekly hours spent in lesson and resource preparation	214 (18)	2.46 (2.32, 2.61)	202 (30)	3.33 (3.15, 3.51)	416 (48)	-0.22 (-0.34, -0.09)	0.76 (0.65, 0.89)	0.001
Complete case analysis	Weekly hours spent in lesson and resource preparation	162 (70)	2.34 (2.19, 2.5)	153 (79)	3.2 (3.01, 3.39)	315 (149)	-0.23 (-0.38, -0.08)	0.74 (0.61, 0.9)	0.002
Primary analysis, RQ1	Weekly hours spent in lesson and resource preparation	214 (18)	2.46 (2.32, 2.61)	202 (30)	3.33 (3.15, 3.51)	416 (48)	-0.22 (-0.35, -0.09)	0.76 (0.64, 0.89)	0.001

Subgroup analyses

RQ5a: What is the impact of Aila on the average time spent per week on lesson planning for KS2 lessons among the subgroup of teachers who indicate low/no prior use of GenAI in lesson planning?

We performed subgroup analyses to explore the differential effect of Aila on hours spent in lesson and resource preparation across the four subgroups based on self-reported prior use of GenAI: those that responded to the baseline survey to say they used GenAI to prepare ‘most/all lessons’, ‘at least one lesson a week’, ‘less often’, or ‘I have never/I no longer use GenAI to support my lesson planning’. As detailed in the Methods section, two approaches were used to explore this effect: an interaction model and separate subgroup models, reported separately below. As with the primary analysis model, the outcomes in the subgroup analyses models were log₁₀ transformed. Note that the study was powered for the primary research question (RQ1) so these subgroup analyses are likely to be underpowered.

Interaction model

Table 27 presents findings from this analysis. The interaction term was not significant ($p = 0.669$). This means we detected no differential impact of Aila on time spent in lesson and resource preparation considering the four prior use of GenAI categories.

Table 27: Impact of Aila on time spent in lesson and resource preparation in analysed subgroups—interaction model

		Unadjusted means				Effect size			
		Aila group		Non-Aila group					
Outcome	Subgroup	n	Mean (95% CI)	n	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: Most/all lessons	15	2.14 (1.64, 2.78)	9	4.66 (3.74, 5.8)	24	-0.44 (-1.07, 0.19)	0.57 (0.26, 1.27)	0.282
Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: At least one lesson a week	67	2.84 (2.57, 3.14)	69	3.55 (3.24, 3.9)	136	-0.19 (-0.46, 0.07)	0.78 (0.56, 1.09)	0.24
Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: Less often	81	2.39 (2.17, 2.62)	57	2.95 (2.65, 3.29)	138	-0.15 (-0.42, 0.12)	0.82 (0.59, 1.16)	0.497
Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: I have never/I no longer use GenAI to support my lesson planning	51	2.22 (1.96, 2.53)	67	3.3 (3.01, 3.62)	118	-0.28 (-0.56, 0.01)	0.7 (0.49, 1.01)	0.063

Separate models

Table 28 presents findings from the separate subgroup models. When analysed separately, these results show that we detected a significant difference in the hours spent in lesson and resource preparation between the Aila and the Non-Aila group for one subgroup: 'I have never/I no longer use GenAI to support my lesson planning'. This suggests there may be a particular impact of using Aila for teachers that did not use GenAI prior to the trial. The effect was slightly larger (that is, a greater time-saving) than the effect seen in RQ1 (fold change: 0.69 vs 0.76). However, this result should be treated with caution due to the increased risk of a false positives associated with multiple significance tests.

Table 28: Impact of Aila on time spent in lesson and resource preparation in analysed subgroups—separate models

		Unadjusted means				Effect size			
		Aila group		Non-Aila group					
Outcome	Subgroup	n	Mean (95% CI)	n	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: Most/all lessons	15	2.14 (1.64, 2.78)	9	4.66 (3.74, 5.8)	24	-0.3 (-0.77, 0.18)	0.65 (0.33, 1.3)	0.223
Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: At least one lesson a week	67	2.84 (2.57, 3.14)	69	3.55 (3.24, 3.9)	136	-0.19 (-0.4, 0.02)	0.79 (0.61, 1.02)	0.075
Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: Less often	81	2.39 (2.17, 2.62)	57	2.95 (2.65, 3.29)	138	-0.16 (-0.37, 0.06)	0.82 (0.62, 1.08)	0.154

Weekly hours spent in lesson and resource preparation	Prior Use of GenAI: I have never/I no longer use GenAI to support my lesson planning	51	2.22 (1.96, 2.53)	67	3.3 (3.01, 3.62)	118	-0.29 (-0.55, -0.03)	0.69 (0.49, 0.96)	0.028
---	--	----	----------------------	----	---------------------	-----	-------------------------	----------------------	-------

RQ5b: What is the impact of Aila on the average time spent per week on lesson planning for KS2 lessons among the subgroup of teachers who indicate low/no prior use of GenAI in lesson planning and who intend to plan a high proportion of their lessons from scratch?

We performed subgroup analyses to explore the differential effect of Aila on hours spent in lesson and resource preparation across the twelve subgroup combinations of the four subgroups based on prior use of GenAI—‘most/all lessons’, ‘at least one lesson a week’, ‘less often’, and ‘I have never/I no longer use GenAI to support my lesson planning’ and the three subgroups based on planning from scratch—‘5%–40%’, ‘41–60%’, and ‘61–100%’. As detailed in the Methods section, two approaches were used to explore this effect: an interaction model and separate subgroup models. As with the primary analysis model, the outcome in the subgroup analyses models were $\log_{10}$ transformed. The study was powered for the primary research question (RQ1) so subgroup analyses are likely to be underpowered.

Interaction model

Table 29 presents findings from this analysis. The interaction terms were not significant (Randomisation Group * Prior Use of Gen AI, $p = 0.816$; Randomisation Group * From Scratch, $p = 0.731$; Prior Use of Gen AI * From Scratch, $p = 0.755$; and Randomisation Group * Prior Use of Gen AI * From Scratch, $p = 0.140$). This means we detected no differential impact of Aila on time spent in lesson and resource preparation considering the twelve subgroups.

Table 29: Impact of Aila on time spent in lesson and resource preparation in analysed subgroups—interaction model

	Unadjusted means				Effect size			
	Aila group		Non-Aila group					
Subgroup	n	Mean (95% CI)	n	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Prior Use Of Gen AI: Most/all lessons from scratch: 5%-40%	8	1.64 (1.16, 2.31)	3	4.77 (3.52, 6.45)	11	-0.77 (-1.9, 0.35)	0.38 (0.09, 1.55)	0.451
Prior Use Of Gen AI: Most/all lessons from scratch: 41-60%	3	2.04 (1.02, 4.07)	4	3.94 (2.35, 6.62)	7	-0.35 (-1.65, 0.95)	0.64 (0.13, 3.29)	0.999
Prior Use Of Gen AI: Most/all lessons from scratch: 61-100%	4	3.61 (2.35, 5.56)	2	5.63 (4.85, 6.54)	6	-0.03 (-1.46, 1.4)	0.96 (0.16, 5.82)	1.000
Prior Use Of Gen AI: At least one lesson a week from scratch: 5%-40%	46	2.74 (2.44, 3.07)	46	3.34 (2.98, 3.74)	92	-0.2 (-0.55, 0.16)	0.78 (0.5, 1.23)	0.780
Prior Use Of Gen Ai: At least one lesson a week From Scratch: 41-60%	14	2.42 (1.87, 3.13)	16	4.01 (3.3, 4.88)	30	-0.29 (-0.91, 0.32)	0.69 (0.32, 1.5)	0.902
Prior Use Of Gen AI: At least one lesson a week from Scratch: 61-100%	7	4.95 (3.83, 6.4)	7	4.1 (3.15, 5.34)	14	0.05 (-0.84, 0.95)	1.07 (0.35, 3.28)	1.000

Prior Use Of Gen AI: Less often from scratch: 5%-40%	54	2.6 (2.35, 2.87)	33	2.59 (2.24, 3)	87	-0.01 (-0.39, 0.37)	0.99 (0.61, 1.6)	1.000
Prior Use Of Gen AI: Less often from scratch: 41-60%	14	2.33 (1.75, 3.09)	12	3.63 (2.85, 4.62)	26	-0.25 (-0.91, 0.42)	0.73 (0.32, 1.7)	0.984
Prior Use Of Gen AI: Less often from scratch: 61-100%	13	1.72 (1.32, 2.25)	12	3.46 (2.84, 4.22)	25	-0.58 (-1.27, 0.1)	0.48 (0.2, 1.14)	0.169
Prior Use Of Gen AI: I have never/I no longer use GenAI to support my lesson planning from scratch: 5%-40%	35	2.18 (1.86, 2.56)	43	3.29 (2.97, 3.65)	78	-0.32 (-0.71, 0.07)	0.67 (0.41, 1.1)	0.217
Prior Use Of Gen AI: I have never/I no longer use GenAI to support my lesson planning from scratch: 41-60%	6	2.89 (2.28, 3.67)	13	2.42 (1.82, 3.23)	19	0.2 (-0.64, 1.05)	1.29 (0.45, 3.75)	1.000
Prior Use Of Gen AI: I have never/I no longer use GenAI to support my lesson planning from scratch: 61-100%	10	2.07 (1.55, 2.78)	11	4.61 (3.79, 5.61)	21	-0.59 (-1.32, 0.14)	0.47 (0.19, 1.19)	0.215

Separate models

Table 30 presents findings from this analysis. When analysed separately, these results show that we detected a significant difference in the hours spent in lesson and resource preparation between the Aila and the Non-Aila group for three subgroups: 'Prior Use of GenAI: Most/all lessons | From Scratch: 5%-40%', 'Prior Use of GenAI: Less often | From Scratch: 61-100%' and 'Prior Use of GenAI: I have never/I no longer use GenAI to support my lesson planning | From Scratch: 5%-40%'. There does not appear to be an explainable pattern to which subgroups are demonstrating significant between group effects. Given there were no significant interactions and due to the increased family-wise error rate due to multiple comparisons, we suggest interpreting these results with caution.

Table 30: Impact of Aila on time spent in lesson and resource preparation in analysed subgroups—separate models

	Unadjusted means				Effect size			
	Aila group		Non-Aila group					
Subgroup	N	Mean (95% CI)	n	Mean (95% CI)	Total n	Standardised effect size (95% CI)	Fold change (95% CI)	p-value
Prior Use Of Gen AI: Most/all lessons from scratch: 5%-40%	8	1.64 (1.16, 2.31)	3	4.77 (3.52, 6.45)	11	-0.73 (-1.25, -0.21)	0.36 (0.17, 0.75)	0.006
Prior Use Of Gen AI: Most/all lessons from scratch: 41-60%	3	2.04 (1.02, 4.07)	4	3.94 (2.35, 6.62)	7	0 (-1.16, 1.17)	1 (0.13, 7.43)	0.999
Prior Use Of Gen AI: Most/all lessons from scratch: 61-100%	4	3.61 (2.35, 5.56)	2	5.63 (4.85, 6.54)	6	0.15 (-1.13, 1.43)	1.19 (0.27, 5.22)	0.816
Prior Use Of Gen AI: At least one lesson a week from scratch: 5%-40%	46	2.74 (2.44, 3.07)	46	3.34 (2.98, 3.74)	92	-0.2 (-0.41, 0.01)	0.79 (0.61, 1.01)	0.056
Prior Use Of Gen AI: At least one lesson a week from scratch: 41-60%	14	2.42 (1.87, 3.13)	16	4.01 (3.3, 4.88)	30	-0.42 (-0.88, 0.04)	0.56 (0.3, 1.06)	0.074
Prior Use Of Gen AI: At least one lesson a week from scratch: 61-100%	7	4.95 (3.83, 6.4)	7	4.1 (3.15, 5.34)	14	0 (-0.52, 0.52)	1 (0.57, 1.74)	0.994
Prior Use Of Gen AI: Less often from scratch: 5%-40%	54	2.6 (2.35, 2.87)	33	2.59 (2.24, 3)	87	-0.01 (-0.29, 0.27)	0.98 (0.7, 1.38)	0.929
Prior Use Of Gen AI: Less often from scratch: 41-60%	14	2.33 (1.75, 3.09)	12	3.63 (2.85, 4.62)	26	-0.32 (-0.97, 0.34)	0.62 (0.23, 1.67)	0.345
Prior Use Of Gen AI: Less often from scratch: 61-100%	13	1.72 (1.32, 2.25)	12	3.46 (2.84, 4.22)	25	-0.56 (-0.83, -0.29)	0.47 (0.33, 0.68)	<0.001
Prior Use Of Gen AI: I have never/I no longer use GenAI to support my lesson planning from scratch: 5%-40%	35	2.18 (1.86, 2.56)	43	3.29 (2.97, 3.65)	78	-0.33 (-0.61, -0.05)	0.66 (0.46, 0.94)	0.023
Prior Use Of Gen AI: I have never/I no longer use GenAI to support my lesson planning from scratch: 41-60%	6	2.89 (2.28, 3.67)	13	2.42 (1.82, 3.23)	19	0.09 (-0.49, 0.67)	1.13 (0.52, 2.45)	0.765
Prior Use Of Gen AI: I have never/I no longer use GenAI to support my lesson planning from scratch: 61-100%	10	2.07 (1.55, 2.78)	11	4.61 (3.79, 5.61)	21	-0.57 (-1.18, 0.03)	0.47 (0.21, 1.03)	0.061

Implementation and process evaluation results

Summary

Teacher context

Trial teachers had a median teaching experience of ten years, with a quarter (25%) in their first five years of teaching and almost half (44%) having leadership responsibilities. Almost all teachers taught a single KS2 class, including a small proportion (9%) who taught multiple year groups in that class.

Usual practice in lesson planning

Before the trial, the expected median weekly planning time was six hours, with 50% of teachers expecting to plan for four to ten hours. Lesson planning was a concern for most teachers, with almost all (90%) saying that lesson planning was a source of stress and over half (61%) reporting they spent too much time on planning. The low proportion of planning from scratch may have limited teachers' opportunities to use Aila as an average (median) trial teacher was planning 21% to 40% of their lessons from scratch.

About three-quarters (76%) of trial teachers had previously used GenAI to support lesson planning. The extent of this use varied significantly, with the median teacher using GenAI less than once a week. More teachers had tried a general GenAI tool, compared with education-specific AI tools. Common prior uses—including writing model texts with specific features and generating high-volume question banks for repeated practice—are not explicitly supported by Aila.

As no trial teachers were using Aila at baseline, the non-Aila approach was equivalent to 'usual practice' for trial teachers, while the Aila approach contrasts with usual practice.

Adherence

Based on self-reported data, 86% of non-Aila teachers were adherent (did not use Aila during the trial) and 81% of Aila teachers were adherent (used Aila at least once during the trial period). For non-adherent teachers in the Aila group, use of Aila was very low (0.6% of lessons), suggesting minimal threat to the contrast established in the trial. While teachers in both groups were asked not to download additional Oak National Academy KS2 resources during the trial, almost half (47%) of the Aila group and almost a quarter (22%) of the non-Aila group did so.

Use of Aila

The mean use of Aila by the Aila group was for 7% of lessons, compared with 15% use of other GenAI tools, most commonly those such as ChatGPT. Comparison with the non-Aila group suggests that Aila was supplementing rather than replacing other GenAI tools. Use of Aila declined as the trial progressed and more than half of teachers in the Aila group (57%) stopped using it before the end of the trial. Where teachers provided reasons for this, most related to negative perceptions of Aila or preferring a different GenAI tool.

Aila was most commonly used when teachers were planning a lesson from scratch, or among specific elements of a lesson, for example when preparing a quiz. Use of the lesson planning function was much higher than the teaching materials function. The most common subject planned with Aila was English, followed by maths and history.

More than half the teachers found it easy to access Aila (65%) and enter lesson details (65%) and agreed that Aila materials were aligned to the national curriculum (71%) and had accurate subject content (54%). However, teachers had more mixed views about the ease of refining materials (both inside Aila and after downloading) and the relevance of materials to their teaching context.

Perceived impact

Teachers who used Aila had varied perceptions on whether using Aila saved time. However, Aila-assigned teachers were more likely than non-Aila assigned teachers to have their view of lesson planning workload improve during the trial period.

More than half of teachers (66%) who used Aila perceived that lessons planned with Aila were lower quality compared with equivalent lessons before the trial, and that pupil engagement and learning were similar (70% and 79% respectively). Over half of teachers who had used Aila *disagreed* that Aila had influenced their approach to lesson planning (58%), helped engage pupils (55%), or helped teaching be more tailored or personalised to their pupils (58%). This may reflect the high levels of confidence in lesson planning at the start of the trial. Where teachers did perceive positive impact from Aila, they were more likely to intend to use Aila in the subsequent (spring) term.

Moderators

Teachers in the Aila group who reported lower prior pedagogical confidence, subject knowledge for teaching, or years of teaching experience at baseline had significantly lower lesson planning workload during the trial compared with teachers with the same baseline characteristics in the non-Aila group. Teachers in the Aila group with higher self-reported technological confidence at baseline had significantly lower lesson planning workload compared with teachers with the same baseline characteristics in the non-Aila group.

Teacher context

Contextual data about teachers was gathered via the baseline survey (n = 464). Similar proportions of teachers were teaching each Key Stage 2 year group (Year 3: 28%, Year 4: 30%, Year 5: 31%, Year 6: 30%). About one in seven teachers (14%) taught more than one KS2 year group. Most commonly, these teachers taught a single class with mixed year groups (9% of all teachers), other patterns included covering Planning, Preparation and Assessment time (PPA), and subject-specialist teaching. Just over a quarter of teachers (27%) shared responsibility for a class.

There was an extensive range of teaching experience across the group of trial teachers, including a high proportion of very experienced teachers. The median level of teaching experience was ten years, while over a quarter (29%) had taught for at least 15 years. A quarter (25%) had up to five years of teaching experience, including the small proportion (8%) who were early career teachers or trainees. Almost all (99%) held Qualified Teacher Status (QTS). Almost half of participating teachers (44%) held leadership responsibilities in addition to their classroom teaching, for example, as a subject lead and/or senior leader.

Usual practice in Key Stage 2 lesson planning

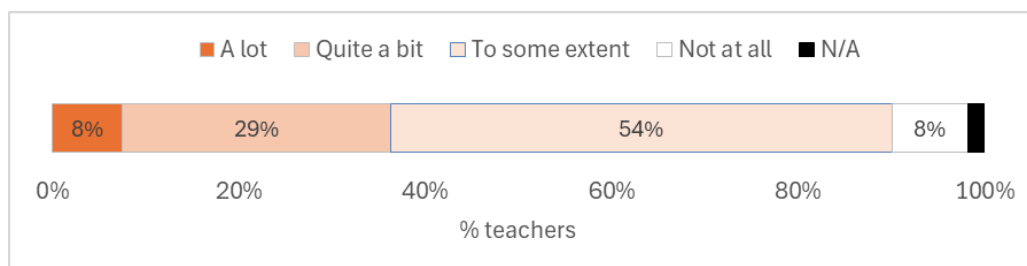
Lesson planning

At baseline, the median teacher expected to teach KS2 for 27 hours per week during the autumn term, with half (50%) reporting contact time of between 20 and 30 hours. The median teacher expected to complete six hours of planning per week for KS2, with half (50%) expecting to plan between four and ten hours. These figures were similar across the Aila and non-Aila groups. These median expected planning hours were higher than the median of planning hours which were subsequently reported via diaries, of four hours per week.

Lesson planning workload was a concern for most teachers

Sixty-one percent of teachers said they had spent too much time on lesson planning in 2024/2025 (including 15% who said 'far too much' and 46% who said 'too much'). Almost all other teachers (36%) said their lesson planning time was 'about right', while very few said they spent 'too little' time on lesson planning (1%) or that the question was not applicable (2%, for example, because they were not teaching in the previous year). Almost all (90%) said lesson planning was a source of stress to at least 'some extent' in their current job, including over a third of teachers (36%) who said lesson planning was a source of stress 'a lot' or 'quite a bit'. See Figure 8 for details.

Figure 8: To what extent is having too much lesson planning to do a source of stress in your work?

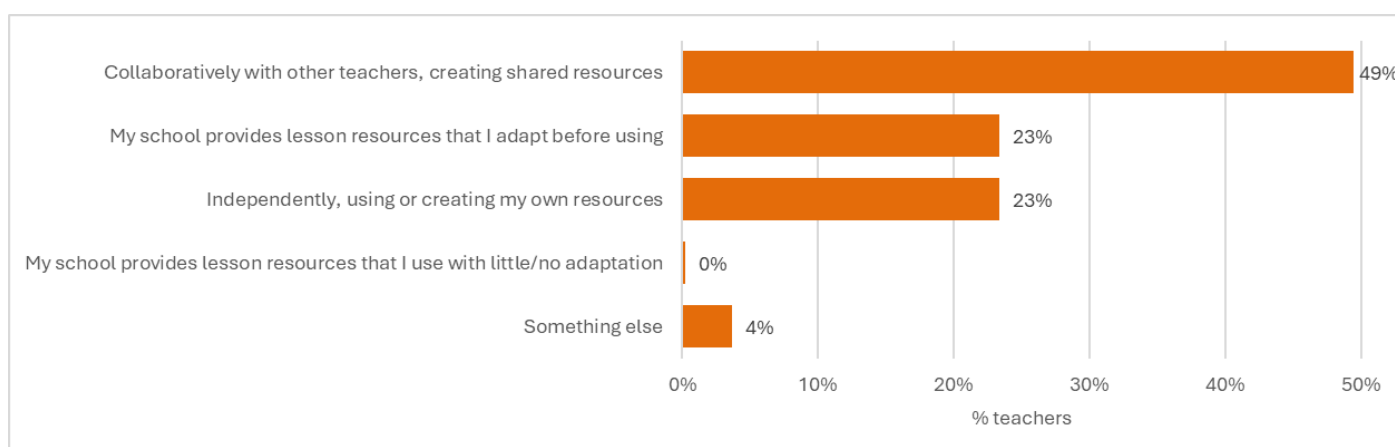


Source: baseline survey, n = 463. Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

Interviewed teachers were asked about their planning rhythm. Teachers were usually aiming to do the bulk of their planning a week in advance and then make adjustments closer to teaching in response to class needs and progress. For teachers who reported average weekly planning of up to four hours, most planning was usually completed as a single block of time, commonly in Planning, Preparation and Assessment (PPA) time, but sometimes after school or at a weekend. Among teachers who reported average weekly planning of more than four hours, planning required multiple blocks of time.

For about half of teachers (49%), collaborative planning (creating shared resources with other teachers) was their most common approach to lesson planning. Almost a quarter of teachers each said their most common approach was adapting school-provided resources (23%) or independent planning using or creating their own resources (23%). See Figure 9 for details.

Figure 9: Think about your lessons for the next week, how were/are **most** of those lessons planned?



Source: baseline survey, n = 463.

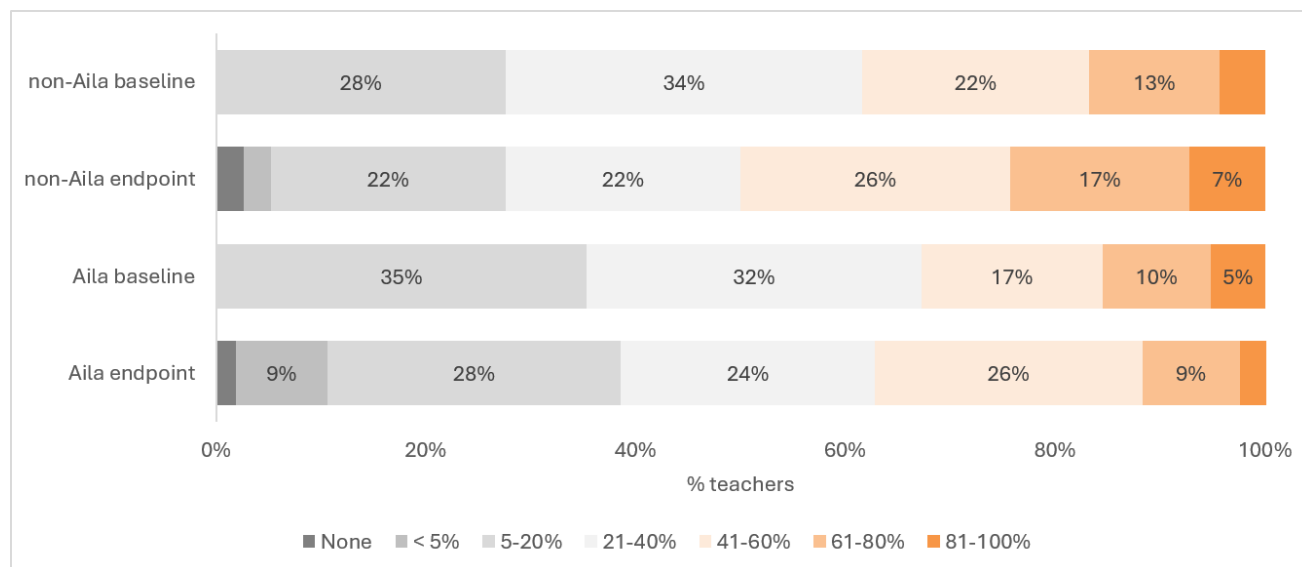
For teachers across both arms who completed both baseline and endpoint surveys, changes were less than ten percentage points for both Aila and non-Aila groups, suggesting there were no substantial changes to these broad approaches to lesson planning during the trial.

The concepts behind these approaches were explored during the interviews. Among interviewed teachers who used collaborative planning, teachers discussed the broad plan for the lesson sequence, and then divided the work, moving to individual planning for either a whole lesson or specific activities or resources. Given that most of this 'collaborative' planning is actually completed individually, this means that Aila, which is designed for individual use, can be used by teachers who describe their lesson planning as 'collaborative'.

At baseline, the average (median) teacher expected to plan 'from scratch' for 21% to 40% of their KS2 planning, which is the main process which Aila models. Proportions of planning 'from scratch' were similar for both the Aila and non-Aila groups. Teachers who expected to plan less than 5% of lessons 'from scratch' were not eligible for the trial and were screened out through the eligibility established through baseline survey. The longitudinal survey showed similar results for planning from

scratch after the trial, suggesting this proportion is stable over time. To explore changes in planning ‘from scratch’ during the trial, we compared a matched sample of teachers who completed both the baseline and endpoint surveys. As shown in Figure 10, the non-Aila group reported that they did more planning ‘from scratch’ in the autumn term (September to December 2025) than they anticipated at baseline. The Aila-assigned group was more similar across baseline and endpoint. *Compared with the logic model, the proportion of planning ‘from scratch’ across trial teachers may act as a limiter for teachers’ opportunity to use Aila.*

Figure 10: Percentage of planning ‘from scratch’—as anticipated at baseline, and as reported retrospectively at endpoint



Source: baseline survey and endpoint survey matched sample, n = 313.

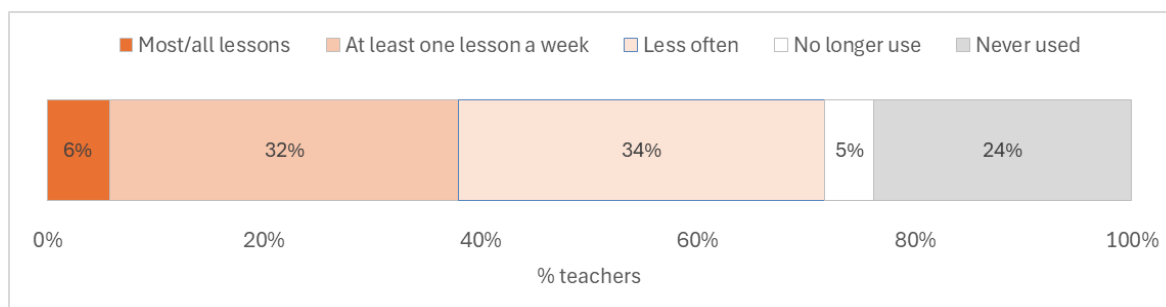
Use of GenAI to support lesson planning

About three-quarters of trial teachers (76%) had used GenAI to support their lesson planning at some point before the trial, while a quarter (24%) had never done so. This proportion who never used GenAI is lower than the equivalent proportion (35%) reported by the Technology in Schools survey (DfE, 2025b), which surveyed teachers a few months before the trial baseline survey (January to May 2025, compared with September 2025). This suggests the extent of any prior use of GenAI was higher for trial teachers than the broader teacher population, which we might expect when participants are volunteering to take part in a trial on a topic they are interested in.

More participating teachers had tried using a general GenAI tool (such as ChatGPT or Microsoft Copilot) to support lesson planning (67%) than an education-specific GenAI tool (24%), while very few trial teachers had ever used Aila (2%), and none were currently using it. Teachers who currently used it more than once per half-term were not eligible for the trial and were screened out through the eligibility established via baseline survey.

The average frequency of use of GenAI was relatively low before the trial, though there was extensive variation. The median teacher used GenAI tools to support lesson planning less than once a week (Figure 11).

Figure 11: Teachers' baseline frequency of use of GenAI tools to support lesson planning



Source: baseline survey, n = 463.

As no trial teachers were currently using Aila for lesson planning, the non-Aila approach was equivalent to 'usual practice' for trial teachers, while the Aila approach contrasts with usual practice.

Based on weekly diary data, most teachers used GenAI tools (other than Aila) to support lesson planning during the trial (91% of Aila-assigned teachers and 75% of non-Aila assigned teachers).¹⁷ Across the ten weeks of the trial, teachers used other GenAI tools to support planning to at least some extent for 15% of the lessons they planned. This was very similar for the two groups (15% Aila, 16% non-Aila) and was similar across each week of the trial (13% to 17%). However, the proportion of use varied across teachers, with the lowest quarter of teachers not using other GenAI at all for planning, and the highest quarter using GenAI for at least 20% of planning.

From endpoint survey data, most teachers reported using a general GenAI tool to support lesson planning during the trial (82% of Aila-assigned teachers and 72% of non-Aila assigned teachers). Relatively few used another education-specific GenAI tool to support lesson planning during the trial (19% of Aila-assigned teachers and 10% of non-Aila assigned teachers).

Interviewed teachers were asked about their usual practice in using GenAI tools (excluding Aila). The most common uses of other GenAI for planning, before and during the trial, were generating example answers (for example, model texts in English which incorporate specific writing features) and generating high volume tailored question banks. These resources do not have stand-alone support in Aila although model texts can be generated as part of a whole lesson within the lesson planning tool.

Beyond lesson planning, more than half of teachers responding to the longitudinal survey reported using GenAI for administrative tasks (60%), while fewer had used it for teaching within a lesson (41%) and only a small proportion had used it for marking (8%).¹⁸ Three-quarters of teachers (75%) reported that their school had provided guidance on staff use of GenAI, most commonly via training (38%), informal guidance (38%), or a formal policy (27%).

Use of Oak National Academy

At baseline, over half of teachers (63%) had used Oak National Academy curriculum resources (33% had not, and 3% did not know). Of teachers who had used Oak National Academy curriculum resources, more than half (59%) were not currently using any in their KS2 teaching, while a further third (36%) were still using them for up to 20% of their lessons. From qualitative data, some teachers who were not currently using Oak National Academy resources were familiar with the website from resources produced during Covid-19 to support remote learning. For example, one teacher was disappointed that the style of Aila-generated resources was different from the remote learning resources, which had included videos of teacher explanations. Compared with the logic model, as less than half of teachers were not currently using Oak National Academy resources for KS2 teaching prior to the trial, this may have acted as a limiter for the motivation to use Aila within the group of trial teachers. It is important to note that trial teachers are likely to be different from the broader population of

¹⁷ The endpoint survey showed similar proportions of teachers using GenAI during the trial (90% of Aila-assigned teachers and 77% of non-Aila assigned teachers). The diary data is the preferred source as it is reported closer to use and has a higher response rate.

¹⁸ Note this question asked about any use of GenAI for these tasks at any point in their career ('ever') rather than just during or since the trial.

KS2 teachers. Teachers who were already using Oak National Academy resources are unlikely to have joined the trial, given teachers were being asked to stop using Oak National Academy for planning during the trial, and frequent users of Aila were not eligible to join the trial.

Teacher adherence to their allocated approaches

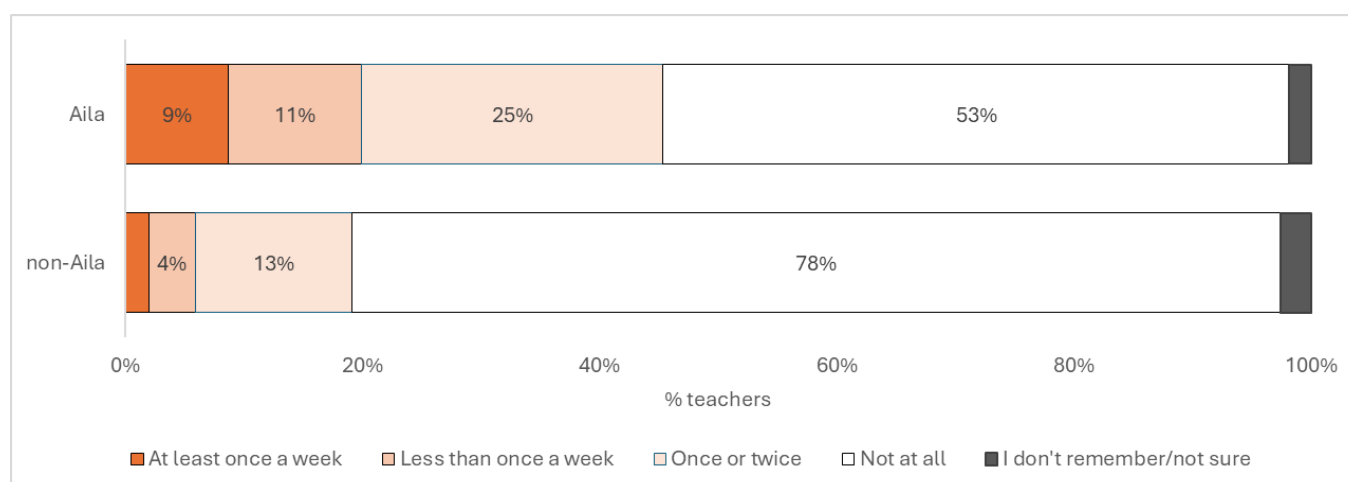
Teachers reported their use of Aila for lesson planning via weekly diaries and the endpoint survey. The diary data is the preferred data source because it had a higher response rate and was filled in weekly, closer to the time of practice. Diary data was used to characterise teacher adherence to the assigned approach.

For non-Aila assigned teachers, overall adherence was high, and the small proportion of non-adherent teachers had low use of Aila. Based on diary data, 86% of the non-Aila group were adherent: they had not used Aila to support any KS2 planning during the trial. Teachers assigned to the non-Aila group who did use Aila (14%) reported relatively low use in their weekly diaries, with a mean use of 0.6% of lessons across the trial. For each single week of the trial, fewer than ten ‘non-Aila’ teachers reported using Aila.

For Aila-assigned teachers, adherence was reasonably high. Based on diary data, 81% of the Aila group were adherent, in using Aila at least once to support KS2 lesson planning. From endpoint survey data, non-adherent teachers generally did not report their reason for not using Aila during the trial. Where teachers did report their reason, the most common responses were that Aila did not meet teachers’ planning needs (n = 5) or that teachers did no planning from scratch (n = 4) during the trial.

Teachers in both groups were asked not to download additional Oak National Academy resources for KS2 teaching during the trial. However, *almost half of Aila-assigned teachers (47%) and almost a quarter of non-Aila assigned teachers (22%) reported using Oak National Academy resources at least once during the trial.* This is higher than the proportion of teachers who were using Oak National Academy resources for KS2 teaching at the start of the trial (22%), suggesting that Aila teachers in particular shifted towards increased use. Most of teachers who used Oak National Academy resources did so ‘once or twice’ across the trial, but 20% of Aila-assigned teachers and 6% of non-Aila assigned teachers used these resources more often (Figure 12). Among teachers who did use Oak National Academy resources, the most commonly cited reasons were previous use of Oak National Academy resources (55%), that is, that use of these resources formed part of teachers’ usual practice, and that Oak National Academy resources covered the content they needed (43%). One interviewed non-Aila teacher had used newly-found Oak National Academy resources: these were found via a general internet search rather than deliberate use of the Oak National Academy website.

Figure 12: Teachers’ use of Oak National Academy resources for new planning during the trial period



Source: endpoint survey, n = 313.

Teacher use of Aila

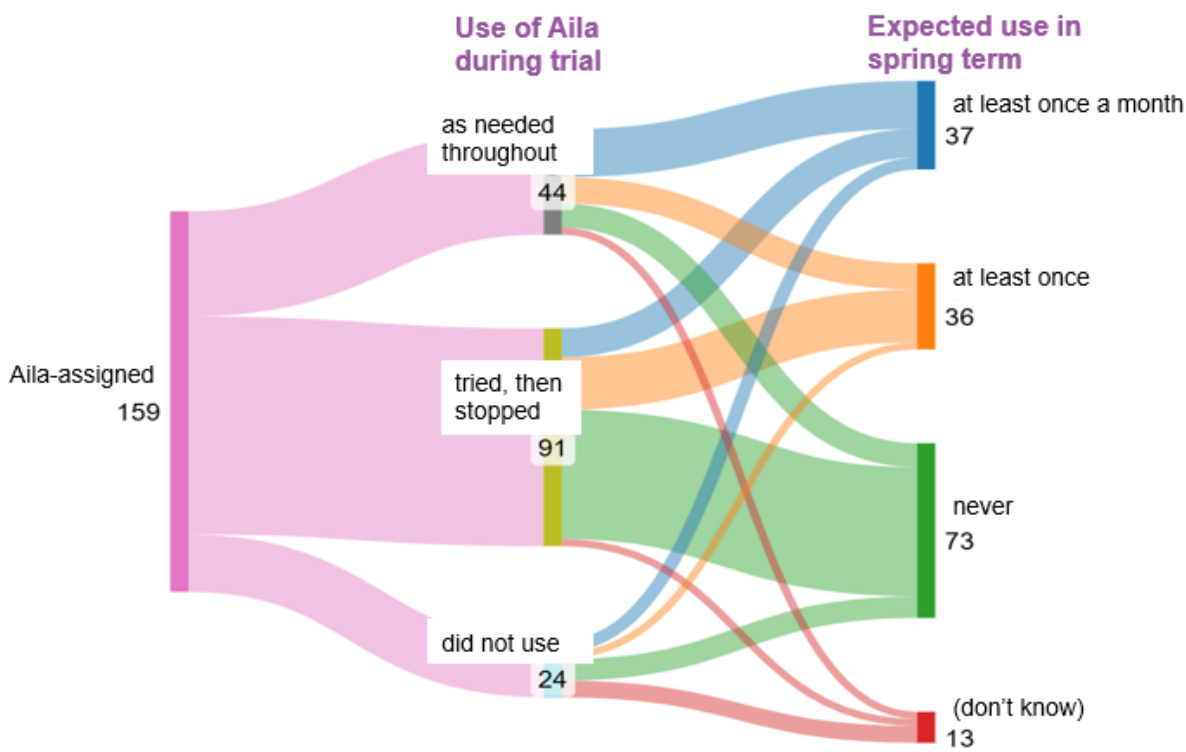
Extent of use

The weekly diary asked teachers to identify what percentage of lessons had been planned with any use of Aila or other GenAI tools, whether to support planning the whole lesson or a small element of it. *For Aila-assigned teachers, the mean proportion of lessons planned using Aila was 7%. Use varied extensively, with a standard deviation of 13%, and the highest quartile of teachers using Aila for planning in at least 10% of lessons.*

For the Aila group, the mean use of Aila per teacher was about half the mean use of other GenAI tools per teacher (15%). The distributions of other GenAI use were similar for the Aila and non-Aila groups. It therefore appears that Aila is primarily supplementing the use of other GenAI tools rather than replacing it. This may reflect the finding that common use cases for GenAI reported by interviewed teachers, such as preparing model texts and question banks (see Usual Practice), do not have stand-alone support within Aila.

The endpoint survey asked Aila-assigned teachers about their use of Aila during the trial and their future intentions. *More than half of Aila-assigned teachers reported that they stopped using Aila during the trial (57%, n = 91), with a further 15% (n = 24) reporting that they did not use Aila at all during the trial.* The remaining 28% kept using Aila during the trial. Intentions to use Aila in the future were higher, with over two-fifths of Aila teachers expecting to use Aila in the spring term (23% each expecting to use it ‘at least once’ and ‘at least once a month’). Patterns of use in the Aila group throughout the trial are illustrated in Figure 13. This shows that Aila-assigned teachers most commonly tried using Aila but then stopped before the end of the trial. Half of teachers who were willing to predict their future use said they would not use Aila in the spring term.

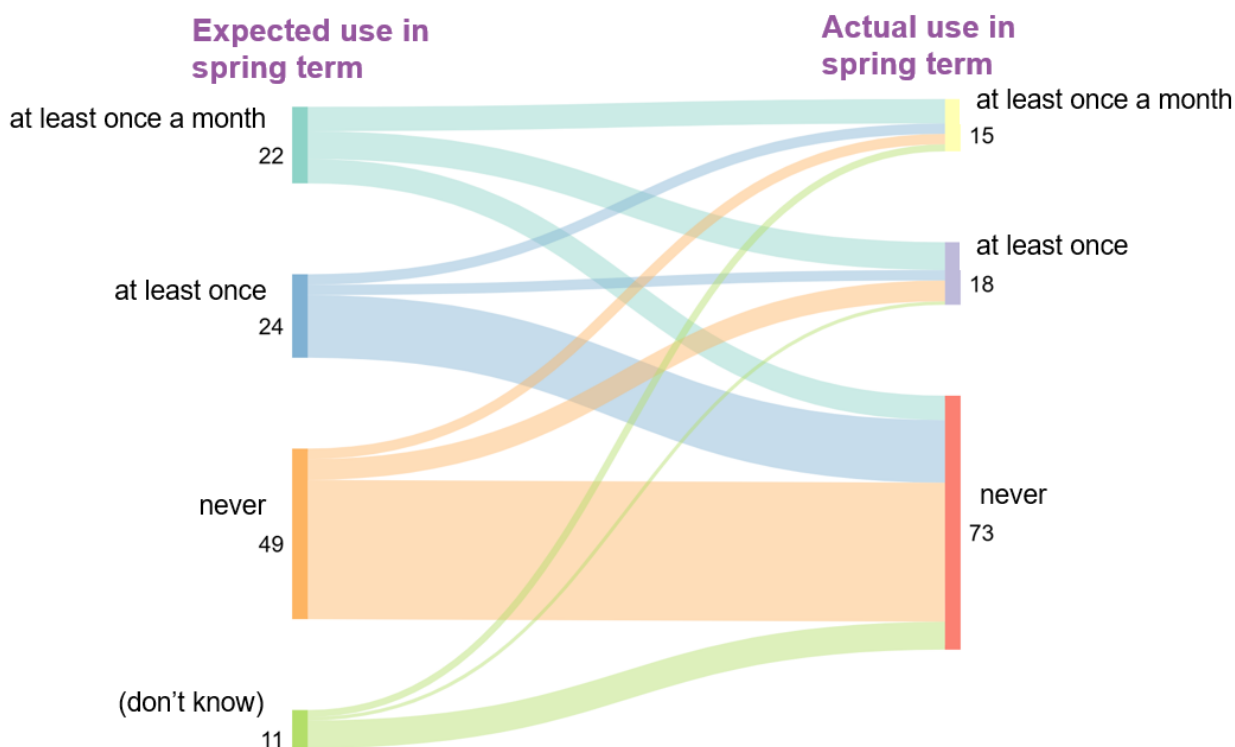
Figure 13: Use of Aila across the trial and future intentions—for Aila-assigned group



Source: endpoint survey, Aila-assigned teachers who characterised their use of Aila during the trial and future intentions, n = 159.

Based on teachers who completed both the endpoint and longitudinal surveys, continuing use of Aila by Aila-assigned teachers (n = 106) in the spring term was slightly lower in practice than teachers anticipated at the end of the trial (Figure 14).

Figure 14: Intended use and actual use of Aila after the trial—for Aila-assigned group



Source: endpoint survey and longitudinal survey, Aila-assigned teachers who characterised their future intentions and reported use, n = 106.

Relatively few non-Aila assigned teachers (9%) used Aila in the spring term, though some reported this was because they were in the control group for the trial and they thought they were supposed to continue not to use Aila (despite trial communications explaining that after the diary period they were no longer required to follow their allocation).

The large proportion of teachers who stopped using Aila during the trial is consistent with the use trend over time for Aila-assigned teachers. *Use of Aila was most prevalent in the first two weeks of the trial and gradually lowered across subsequent weeks*: the proportion of teachers whose weekly planning included Aila declined across the course of the trial from about half in weeks one and two (53% and 54% respectively) to about a third in weeks nine and ten (32% and 33% respectively). In comparison, the proportion of teachers whose weekly planning included other GenAI tools stayed relatively consistent across each week of the trial (59% to 67%). Details are shown in Table 31 and Figure 15.

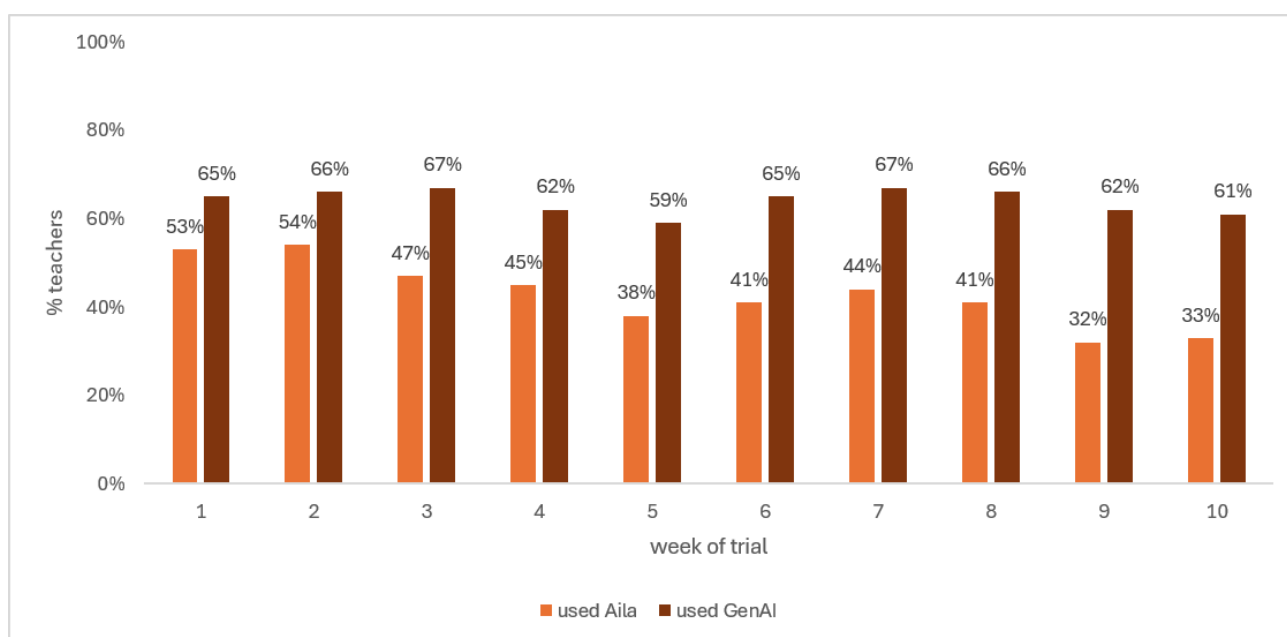
Table 31: Mean and median use of Aila in each week—Aila-assigned teachers

Week	Teachers completing diary who reported >0 hours of planning for that week (n)	% of lessons where Aila was used	
		Mean	Median
1	202	10	5
2	194	10	5
3	189	8	0
4	174	7	0
5	111	7	0
6	175	7	0

7	178	7	0
8	167	7	0
9	155	5	0
10	146	5	0
Overall (weeks 1–10)	212	7	0

Source: diary data, n = 212.

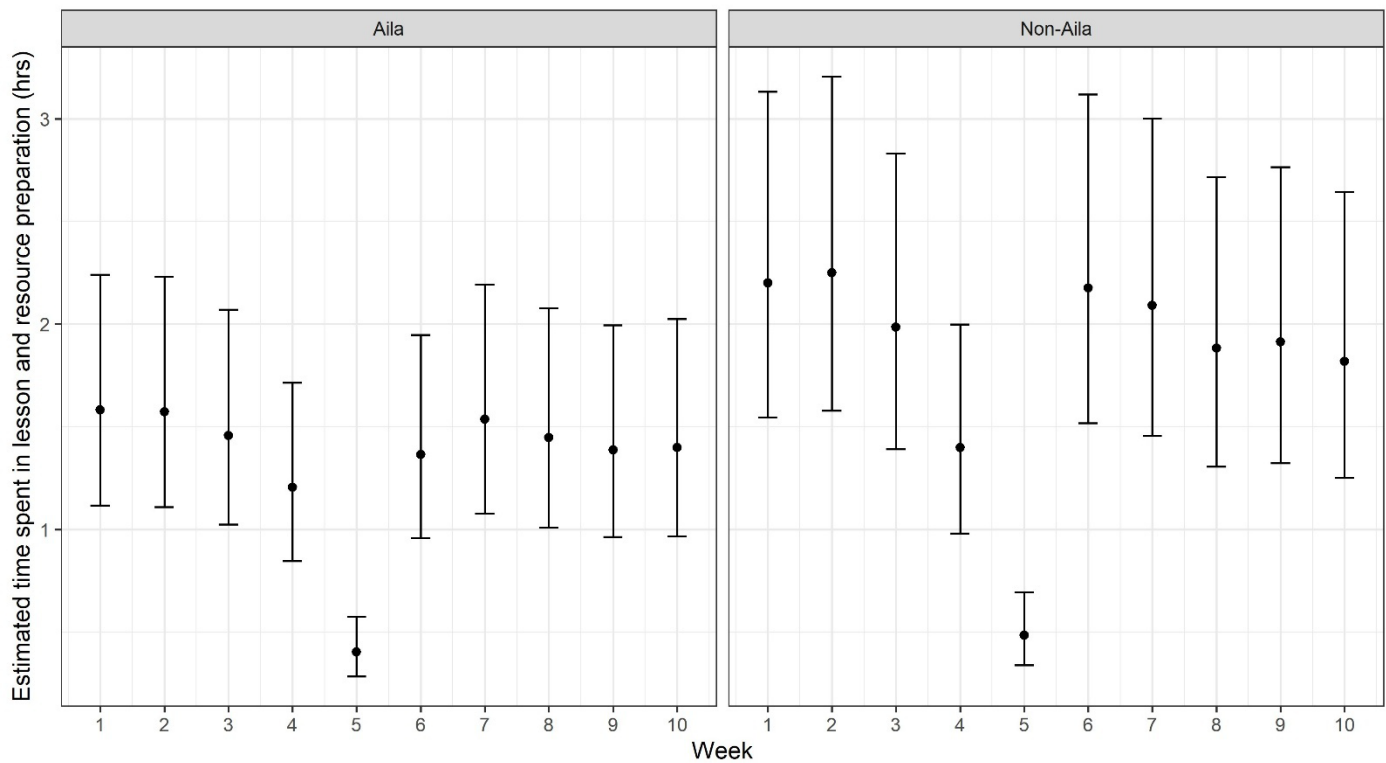
Figure 15: Proportion of Aila-assigned teachers using Aila and other GenAI for lesson planning in each week of the trial



Source: diary data, n = 212.

For both the Aila and non-Aila groups, the weekly lesson planning time remained similar throughout the trial (with the exception of the half-term week), suggesting that the reduced use of Aila did not relate to reduced need for lesson planning. See Figure 16 for details.

Figure 16: Weekly lesson planning time across the trial—Aila group and non-Aila group



Source: diary data, n = 415.

Teachers’ decision-making about using GenAI tools to support lesson planning is explored in a later section.

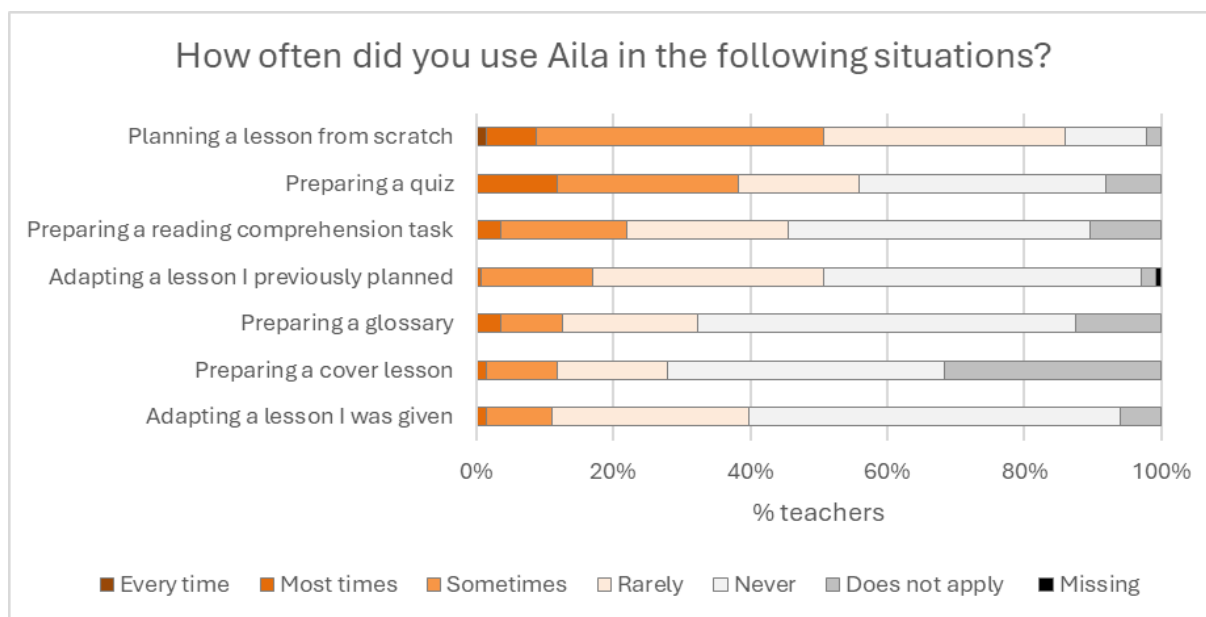
How teachers used Aila

Among teachers who used Aila (n = 136),¹⁹ the most common uses reported in the endpoint survey were for planning a lesson from scratch (51% used Aila at least ‘sometimes’, including 9% ‘most’ or ‘every time’) or preparing a quiz (38% at least ‘sometimes’, including 12% ‘most’ or ‘every time’; see Figure 17). These tasks align with the intended functionality of Aila. Tasks which Aila intends to support which were less frequently used included preparing a reading comprehension task (used at least ‘sometimes’ by 22%) and preparing a glossary (used at least ‘sometimes’ by 13%).

Most teachers that had used Aila reported refining Aila’s generated materials to at least ‘some’ extent, either within the Aila planning process or after download. Aila pauses during the lesson planning process to encourage teachers to make refinements, and resources are designed to be editable after download to allow for further adaptations. For adaptations ‘within the Aila process’, 23% of teachers reported refining materials ‘to a great extent’, while a further 50% reported doing this ‘to some extent’. For adaptations after downloading resources, 29% of teachers reported refining materials ‘to a great extent’, while a further 49% reported doing this ‘to some extent’. Only around one in ten teachers reported not refining the materials at all. See Figure 18 for details.

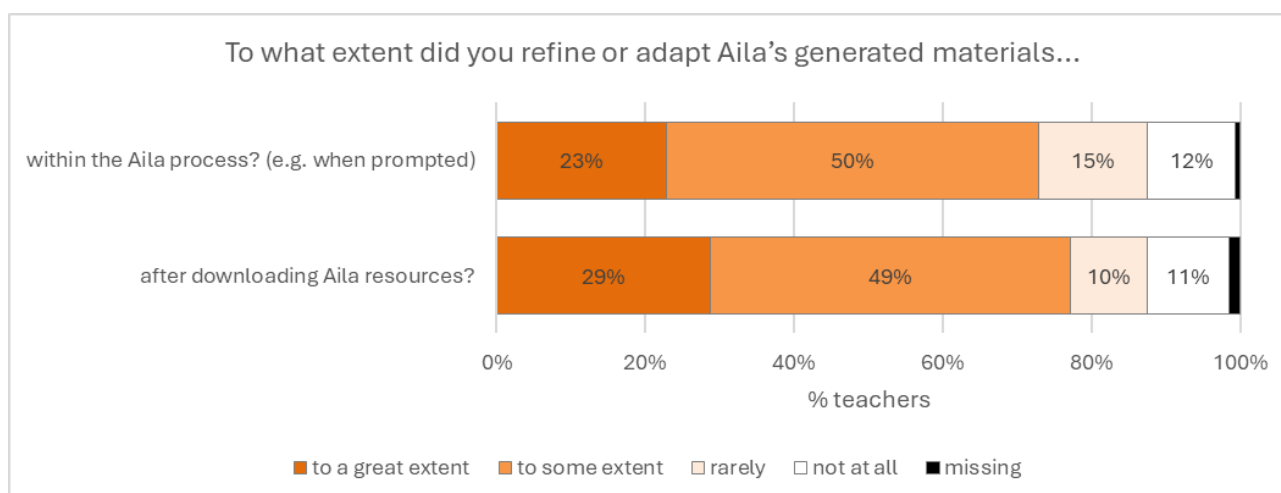
¹⁹ Throughout this section, the phrase ‘teachers who used Aila’ refers to teachers in the Aila group who reported in the endpoint survey that they used Aila during the trial, (i.e. compliant Aila teachers). These analyses exclude the small proportion of teachers in the non-Aila group who used Aila. The phrase ‘Aila-assigned teachers’ and ‘Aila group’ refers to all teachers assigned to the Aila group (i.e. intention to treat), whether or not they used Aila during the trial.

Figure 17: Teacher use cases



Source: Endpoint survey, Aila-assigned teachers who reported using Aila, n = 136.

Figure 18: Extent of teachers' use of refinements



Source: endpoint survey, Aila-assigned teachers who reported using Aila, n = 136.

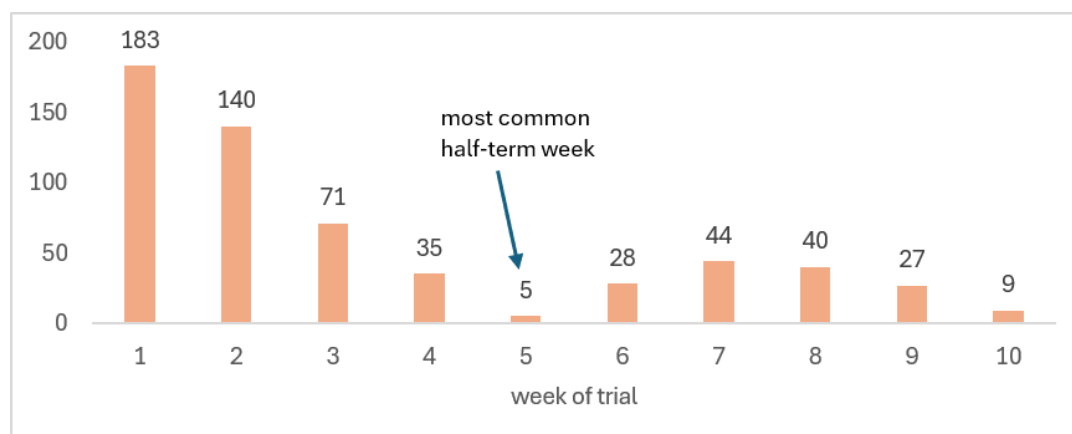
Oak National Academy usage data

As an additional data source to triangulate against, Oak National Academy shared anonymised teacher usage data, matched via teachers' trial contact email or an alternative log in email provided via the endpoint survey. While the usage data points to a similar pattern of behaviour as the self-report in the diaries and surveys, it appears that the usage data is not complete, which is likely to be due to missing login/email details which prevented a full match to the Academy dataset (for example, because teachers may have used multiple emails to log in or not provided an alternative login email via the endpoint survey) or because teachers did not accept cookies on the Aila webpages so their usage could not be tracked. We recommend this issue is considered in advance by any similar trials in the future. For this reason, we recommend caution around the data presented from the usage dataset, while noting the general trends are supportive of the narrative from the other IPE methods. *The usage data showed that, across the trial, 125 of the Aila-assigned teachers planned 582 lessons in*

the ‘lesson planning’ function.²⁰ The median number of uses per teacher was three. The ‘teaching materials’ function (used to create standalone resources) was used relatively rarely (55 times by 25 Aila-assigned teachers) and was not used after the first three weeks of the trial, suggesting that teachers moved away from use of this function. Qualitative interviews suggested that many teachers had not noticed the ‘teaching materials’ function.

The Oak National Academy usage data (Figure 19) shows a similar reduction in use across the course of the trial to that reported by teachers in the diary, with the highest use in the first two weeks.

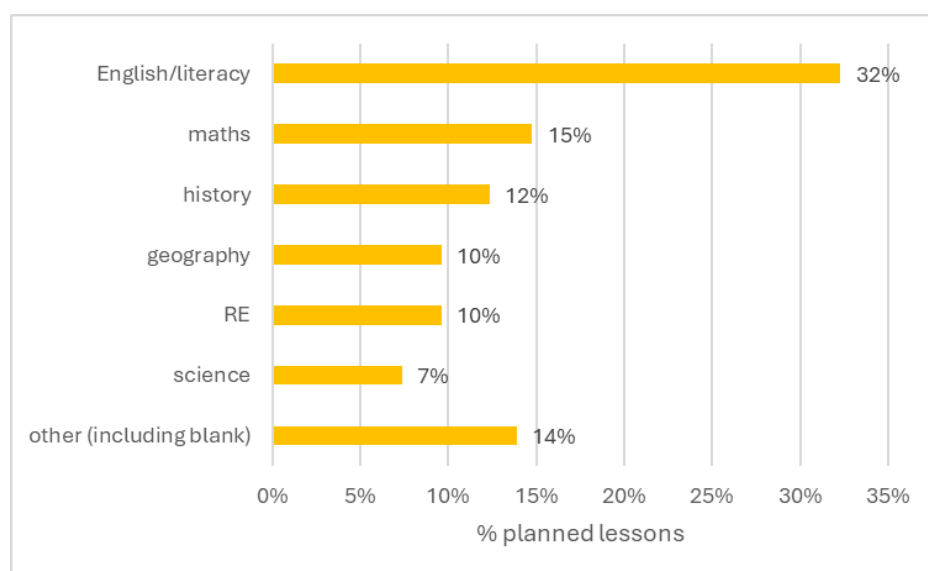
Figure 19: Number of lesson plans started in Aila across the trial (from usage data)



Source: Oak National Academy’s Aila usage data.

Based on Oak National Academy usage data (n = 582 lessons), the most common subject planned within Aila’s lesson planning function was English/literacy (32%), followed by maths (15%), history (12%), RE (10%), and science (7%), with other named subjects accounting for 12% of lessons and 2% where the subject was not entered (Figure 20).

Figure 20: Proportion of lessons planned from scratch in each subject according to Oak National Academy usage data



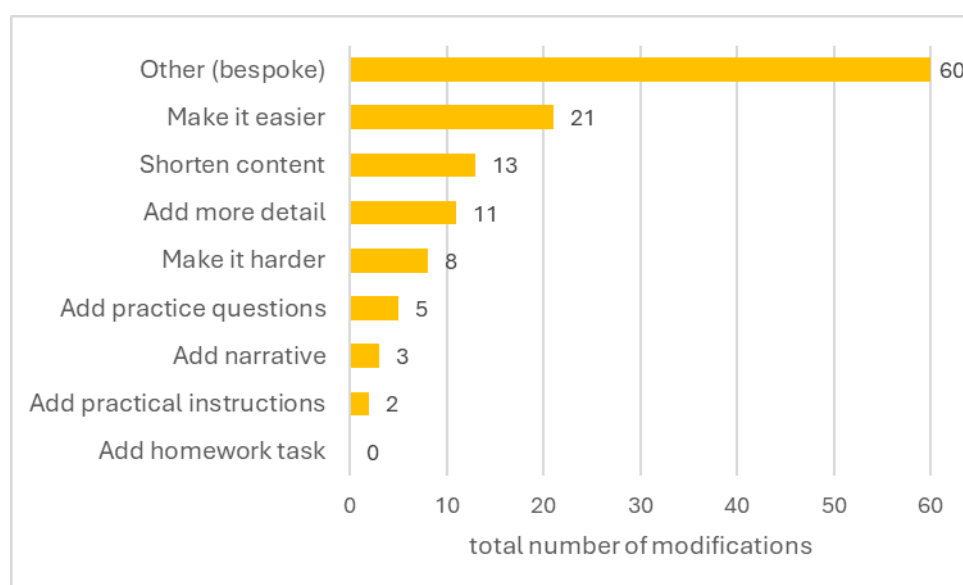
Source: Oak National Academy usage data, lessons planned by Aila-assigned teachers, n = 582.

²⁰ Usage data has been analysed for all lessons planned by Aila-assigned teachers in the trial, excluding three lessons planned by non-Aila assigned teachers. While most of these lessons were labelled by teachers as Key Stage 2 (83%), a small proportion of lessons was left blank (10%) or labelled by teachers as another key stage (6%). We chose to include all lessons in analysis, irrespective of their Key Stage label, because we cannot be sure whether the lesson was used with Key Stage 2 pupils (for example, to revisit Key Stage 1 content).

In the ‘teaching materials’ function, comprehension tasks accounted for almost half of the resources generated (n = 21), of which most were for English (n = 12). Starter or exit quizzes were most commonly generated for maths (n = 9 of 17), while glossaries were most commonly generated for history (n = 6 of 8). From qualitative data, some teachers were using Aila’s lesson planning function to generate starter and exit quizzes rather than using the stand-alone teaching materials function, suggesting low awareness of the stand-alone functions.

Oak National Academy usage data suggested that most lessons planned within Aila’s lesson planning function had no modifications made within Aila (88%), with a further 6% of lessons having a single modification. Considering the different types of modifications available, about half of modifications were bespoke requests (n = 60 of 132), while the most commonly used ‘standard’ modifications were ‘make it easier’ (n = 21), ‘shorten content’ (n = 13) and ‘add more detail’ (n = 11). See Figure 21 for details. Within Aila’s teaching material function, there was a range of one to four modifications, with two being the most common number of modifications.

Figure 21: Teacher use of modifications within Aila

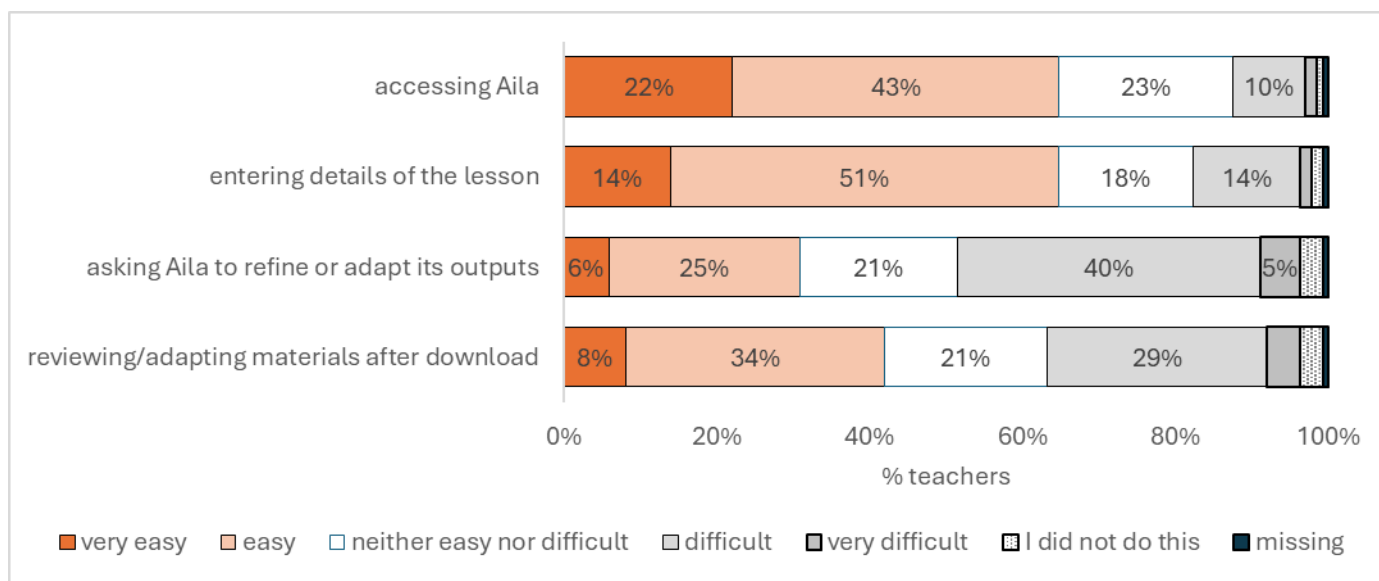


Source: Oak National Academy usage data, lessons planned by Aila-assigned teachers, n = 582.

About half of teachers (51%) agreed that Aila was easy to use, while about a third (29%) disagreed. Teachers were also asked about the ease or difficulty of specific aspects of using Aila (Figure 22). More than half of teachers found it easy to access Aila (65% ‘easy/very easy’, 11% ‘difficult/very difficult’) and enter the details of the lesson (65% ‘easy/very easy’, 15% ‘difficult/very difficult’). Teachers had more mixed views about asking Aila to refine or adapt outputs and about reviewing and adapting materials after downloading them. Similarly, about half of teachers (49%) disagreed that ‘it was quick to make Aila’s outputs ready for use’, while less than a quarter (21%) agreed.

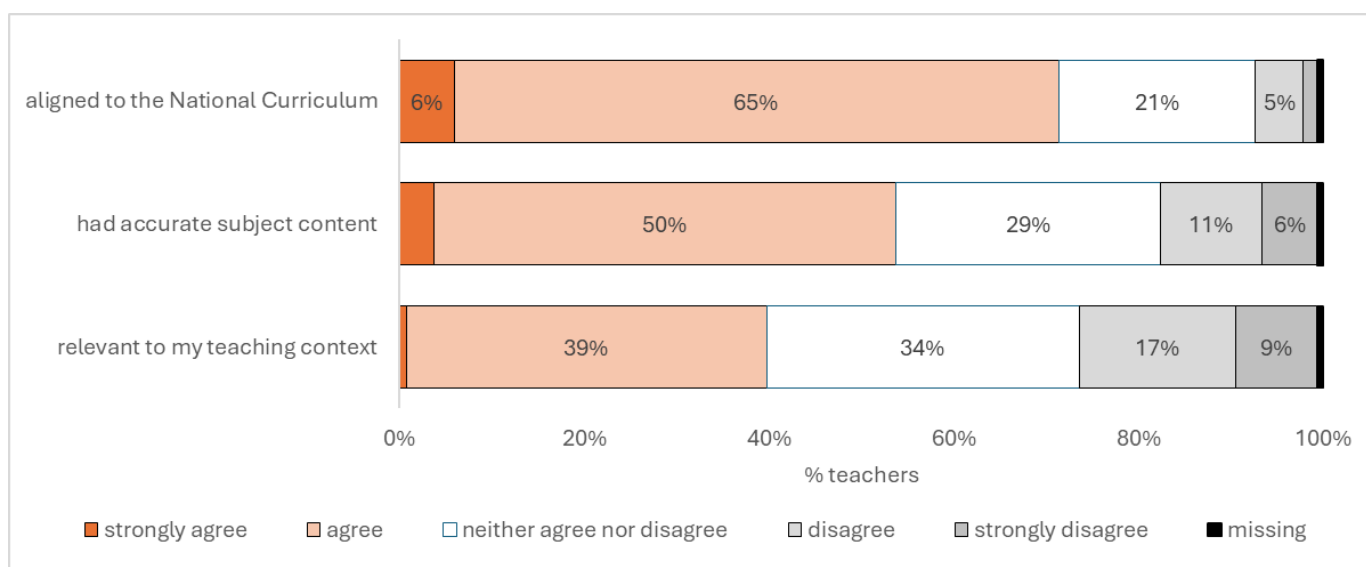
More than half of teachers agreed that Aila-generated materials were aligned to the national curriculum (71%) and had accurate subject content (54%). However fewer (40%) agreed that Aila-generated materials were relevant to their teaching context; 34% had mixed views and 26% disagreed with this statement. See Figure 23 for details.

Figure 22: Ease of using Aila-Teachers' views



Source: endpoint survey, Aila-assigned teachers who reported using Aila, n = 136.

Figure 23: Teachers' views of Aila-generated materials

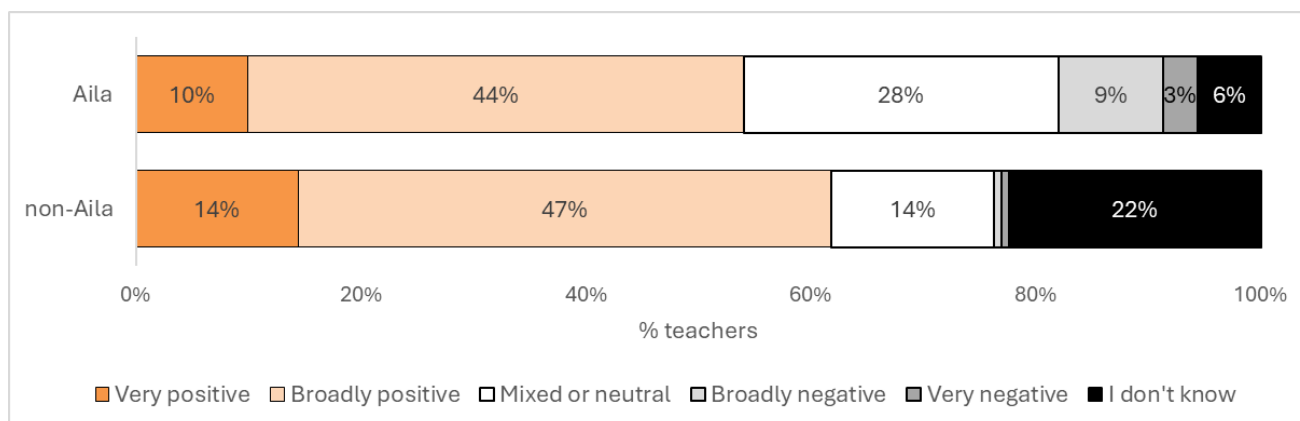


Source: endpoint survey, Aila-assigned teachers who reported using Aila, n = 136.

Teachers' decisions on using Aila and other GenAI tools

At the end of the trial, *there were mixed views across the population of trial teachers about using GenAI to support lesson planning*. Over half of teachers held very/broadly positive views (54% of Aila-assigned teachers and 62% of non-Aila assigned teachers), while over a third held mixed or neutral views, or didn't know (34% of Aila-assigned teachers and 37% of non-Aila assigned teachers). Only a small proportion of teachers held very/broadly negative views (12% of Aila-assigned teachers and 1% of non-Aila assigned teachers). See Figure 24 for details.

Figure 24: Teachers' views on using GenAI to support lesson planning—at the end of the trial



Source: endpoint survey, n = 313.

Reasons for stopping use of Aila

Aila-assigned teachers who reported that they stopped using Aila during the trial were asked the reasons why. Among those who responded to this open-text question (n = 57), four main overlapping reasons were mentioned. Some (n = 9) reported they simply did not need to use Aila because they were not completing substantial planning—for example, because their teaching load had changed—or because they already had most planning in place and the planning functions they needed—such as question banks (see also Usual Practice)—were not supported by Aila.

However, the majority of teachers who stopped using Aila and gave their reason for doing so reported negative perceptions of Aila or preferring a different GenAI tool. Some (n = 26) described interactions with Aila as ‘slow’, ‘inflexible’, ‘requiring lots of steps’, or ‘difficult’. Some teachers (n = 40) had concerns about the outputs generated by Aila. The most common concern was that the resources as generated were not engaging or varied enough for KS2 teaching, including the absence of images and the similar structure. These teachers described the resources as ‘tedious’, ‘dull’, or ‘formulaic’. Other concerns about Aila resources included that Aila did not match the school or teachers’ lesson structure or teaching style, lack of detail in resources, slide decks that were too long for lesson timings, and the absence of tailoring to a specific class. Reflecting their concerns about Aila resources as generated, teachers also reported the need to spend a lot of time adapting resources as a reason they stopped using it.

Some teachers (n = 17) found that a different GenAI tool better fitted their needs. These teachers described other GenAI tools as more flexible, easier to interact with, quicker to use, and able to generate resources that were perceived to be higher quality, more finalised, or more engaging. This included comparisons to general GenAI tools as well as education-specific tools. Based on longitudinal survey data (n = 192), in spring term 2026 the most commonly used GenAI tools for lesson planning were a general GenAI tool (used by 93% of teachers), TeachMateAI (25%), and Canva Magic Studio (17%).

Very few teachers that had stopped using Aila mentioned concerns about Aila’s alignment to the national curriculum or accuracy of subject content. This suggests that teachers’ decisions are more influenced by other factors, particularly planning flexibility, variation, and time. Some teachers who had stopped using Aila also commented positively on some features (n = 5), most commonly the quizzes (n = 3) that they had chosen to use in lessons.

In the longitudinal survey, teachers who had rarely or never used GenAI for lesson planning in spring term (January to March 2026) were asked why (n = 60). Most of the common reasons overlapped with those mentioned for Aila above (not needing to complete substantive planning, lack of images or engaging resources, finding it quicker or better to plan for themselves, and resources needing significant adaptation). Another common reason was not feeling confident or knowing how to plan with GenAI. A few teachers mentioned schools not allowing use of GenAI to plan, while one teacher mentioned not using GenAI tools for environmental reasons. When teachers who never or rarely used GenAI (n = 76) were asked if anything would encourage them to use it for planning, about half each said ‘yes’ (51%) and ‘no’ (49%). The most common requests were for training and for GenAI tools that saved more time.

Interviewed teachers who had stopped using Aila during the trial, or preferred a different GenAI tool, expressed similar concerns to surveyed teachers. Two representative examples from the range of interviewed teachers are provided below.

Teacher A had substantive prior experience using GenAI for planning, including developing his own bespoke ‘agent’ for Microsoft Copilot, which ‘knows the style of maths that we follow at school, the way that our English lessons are formed’ and follows the preferred style and structure. He also used GenAI to review his own lesson ideas: ‘I’ll put my idea for a lesson in [and ask] ... does this hit national curriculum objectives? Are there any other examples I could use?’

Compared with his previous GenAI approach, he found Aila lacked customisation: ‘It can be a little bit robotic ... it’s always the same. It’s always the quiz, it’s always that little explanation.’ He also noted that it constrained teachers’ inputs: ‘There wasn’t that conversational back and forth that you can have with other kind of tools.’ As Aila did not have functionality to create a whole lesson sequence as a planning overview, he used another GenAI tool to do this before swapping to Aila.

Teacher A did not intend to continue using Aila after the trial, preferring to use his bespoke Microsoft Copilot agent: ‘If I’m choosing to use another AI to help me use [Aila], I feel like that’s two steps I don’t need to make.’

Teacher B had not previously used GenAI to plan lessons. She tried using Aila for geography and history lessons. While she felt that Aila provided a useful outline, quizzes, and glossaries, ‘it doesn’t actually give you resources’ to engage pupils. For example, in a history lesson on Boudicca, Teacher B instead used a BBC Schools Radio dramatization and maps of the battlefields to engage pupils: ‘For them to sit and just imagine and be there, I thought it was quite powerful. And they’ve really, really enjoyed it.’

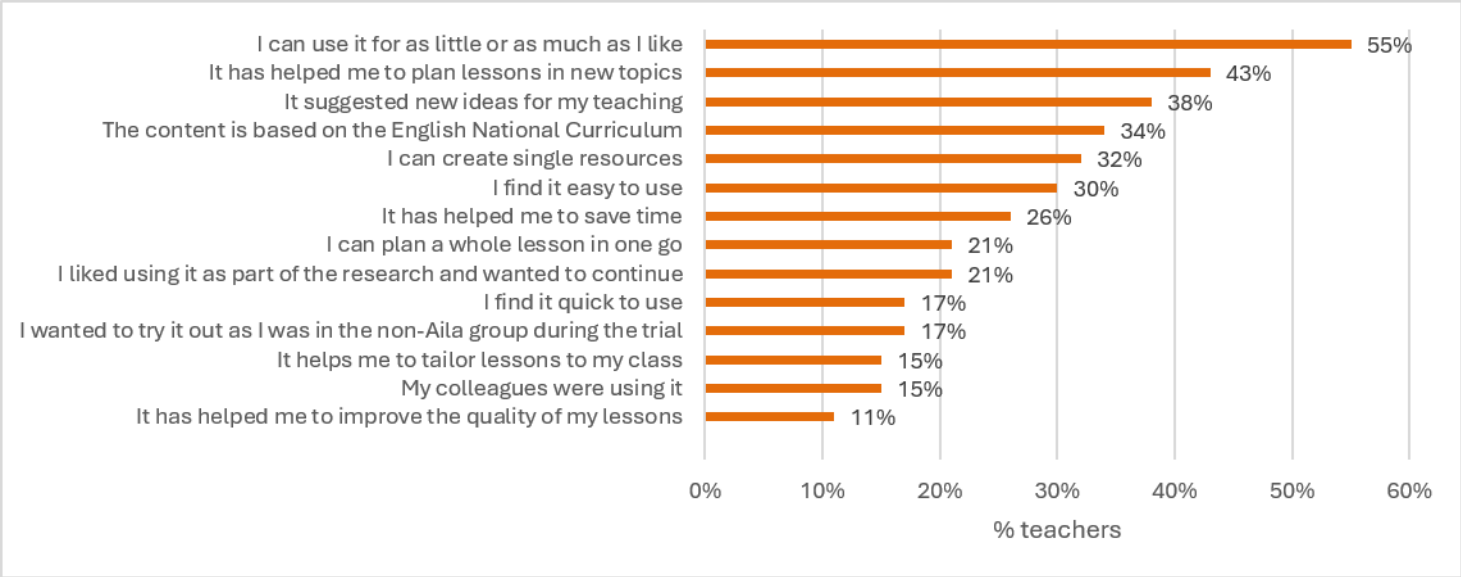
Teacher B prefers to plan lessons and find resources herself rather than delegating this role to a GenAI tool. ‘I really feel like you need a depth of understanding of what you’re teaching. And so I don’t think it’s appropriate to just go on [to a GenAI tool] and then put that on slides.’ She did not intend to continue using Aila after the trial.

Based on longitudinal survey data, the most common reasons teachers gave for using Aila after the trial (spring term 2026, n = 47) were because ‘I can use it for as little or as much as I like’ (55%), ‘it has helped me plan lessons in new topics’ (43%), ‘it suggested new ideas for teachers’ (38%), and ‘the content is based on the English national curriculum’ (34%). See Figure 25 for details. The most common reasons teachers gave for using other GenAI tools after the trial (spring term 2026, n = 191) were because ‘it has helped me save time’ (79%), ‘I find it easy to use’ (74%), and ‘I find it quick to use’ (74%). See Figure 26 for details. Again, this suggests that teachers’ decision to use other GenAI tools are driven by perceptions of saving time and speed and ease of use, which are less commonly reported as reasons by teachers who used Aila.

Where interviewed teachers were using both Aila and another GenAI tool, we asked about how they used each tool. Most commonly, teachers described using different tools for different functionalities. They described preferring general GenAI tools (commonly ChatGPT or Microsoft Copilot) over Aila for ‘quick’ generation, for example providing a single slide or writing model. Teachers also reported that general GenAI tools were more ‘conversational’ than Aila, or that Aila required more upfront information than other tools before generating a resource. For some teachers, a general GenAI tool was built into their computer, which made it quicker to access compared with Aila.

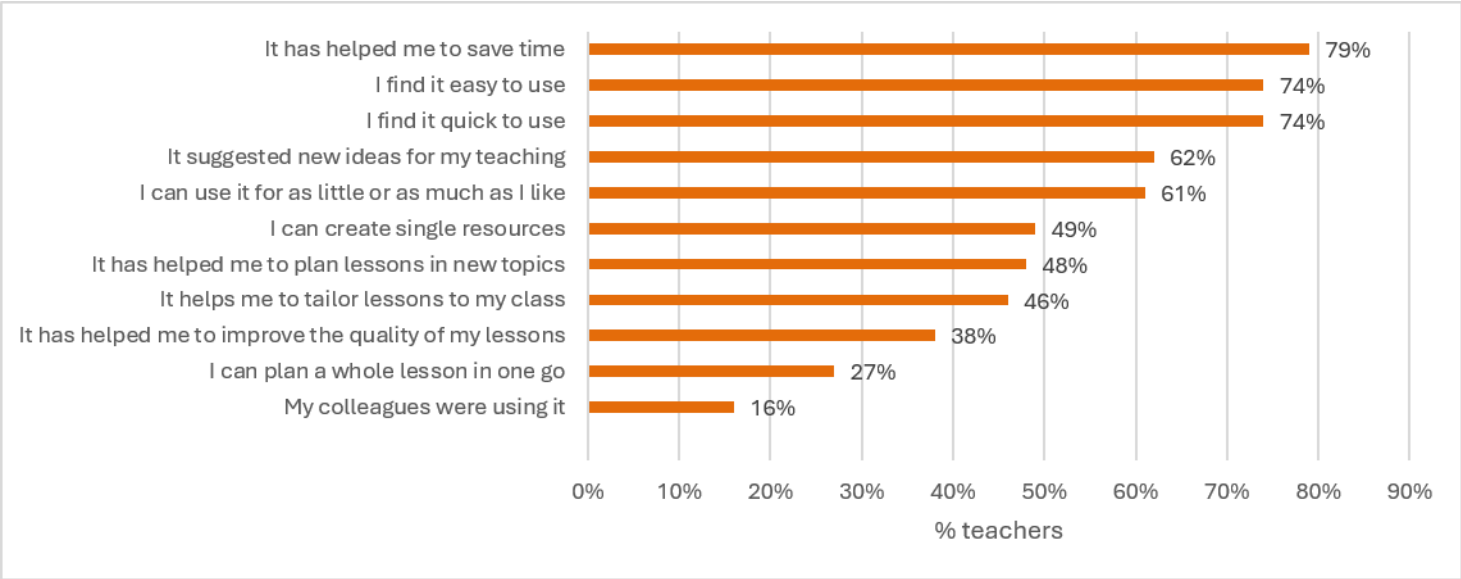
Teachers reported that Aila was more appropriate for creating overarching lessons with good pedagogical structure, for example, aligning with the National Curriculum, checking prior knowledge, breaking down the learning ‘into chunks’, and pauses to check understanding, such as ‘mini recaps’ and ‘exit tickets’.

Figure 25: Teachers’ reasons for using Aila after the trial (spring term 2026)



Source: longitudinal survey, all teachers who used Aila, n = 47.

Figure 26: Teachers’ reasons for using other GenAI tools after the trial—spring term 2026



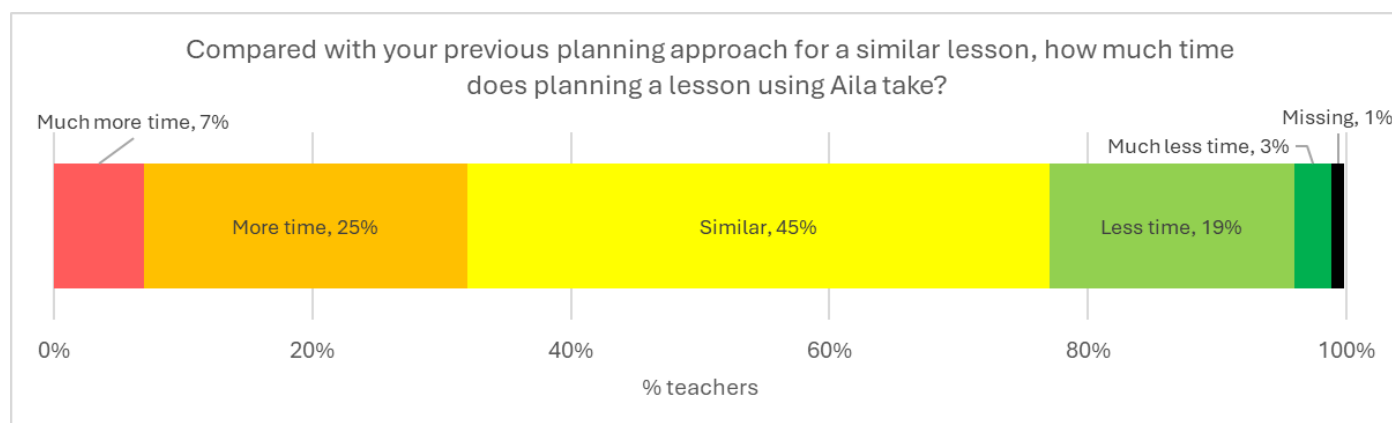
Source: longitudinal survey, all teachers who used other GenAI tools, n = 191.

Perceived impact of Aila

Perceived impact on lesson planning workload

Teachers who used Aila²¹ had varied perceptions on whether using it saved time, compared with similar lessons planned using a different approach. The most common response was that the time spent when using Aila was ‘similar’ (45%), while about a third (32%) perceived that planning with Aila took more time. Less than a quarter of teachers (22%) perceived that planning with Aila took less time. Please see the Conclusion for a discussion about how this relates to the impact results.

Figure 27: Perceived impact on the time taken to plan a lesson



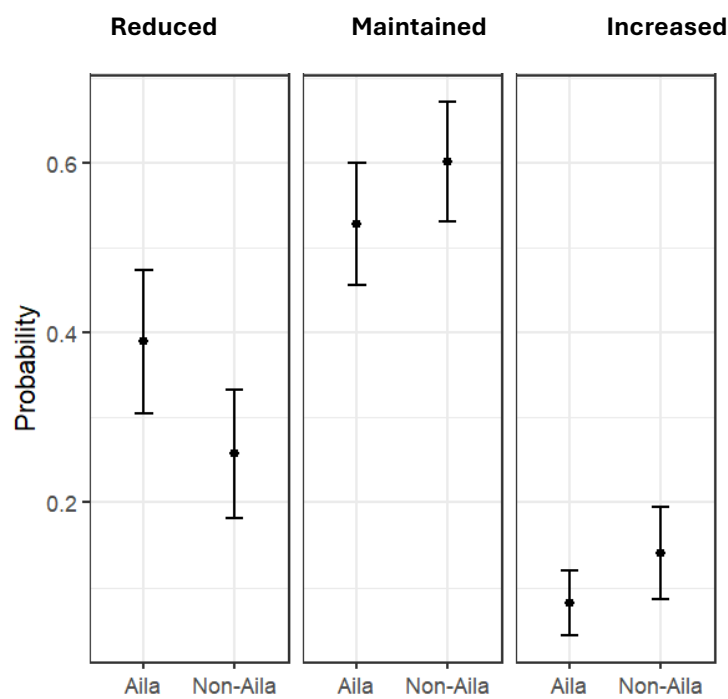
Source: endpoint survey, Aila-assigned teachers who reported using Aila, n = 136.

Among teachers who perceived that planning a lesson with Aila took less time (n = 31), the most commonly reported uses of this ‘time back’ were for other professional responsibilities, including for other aspects of lesson planning and preparation (60%), other teaching responsibilities (50%), and administration, management, or leadership responsibilities (43%). Fewer teachers (37%) reduced their overall workload. Among those who perceived that planning a lesson with Aila took more time (n = 45), the most commonly reported reason was the additional time to review and adapt Aila outputs (70%) followed by the reasons that lesson preparation when using Aila included additional activities (48%) or that it took additional time to learn how to use Aila (39%).

Aila-assigned teachers were significantly more likely than non-Aila assigned teachers to become more positive about their lesson planning workload during the trial period. Teachers in both groups (n = 300) responded to the question ‘Was the amount of time you spent on lesson planning too little, too much or about right?’ in both baseline and endpoint surveys, using response options ‘far too much’, ‘too much’, ‘about right’, ‘too little’, and ‘far too little’. Analysis categorised the response patterns as ‘reduced’, ‘maintained’, or ‘increased’ perceived workload during the trial by comparing baseline to endpoint responses. The most common category for teachers in both groups was that their view of their lesson planning workload was ‘maintained’ (that is, they responded in the same way at both timepoints). After controlling for baseline lesson planning workload, teachers in the Aila group had a significantly higher likelihood than non-Aila teachers to perceive that their lesson planning workload had ‘reduced’ (Figure 28).

²¹Throughout this section, the phrase ‘teachers who used Aila’ refers to teachers in the Aila group who reported in the endpoint survey that they used Aila during the trial, (i.e. compliant Aila teachers). These analyses exclude the small proportion of teachers in the non-Aila group who used Aila. The phrase ‘Aila-assigned teachers’ and ‘Aila group’ refers to all teachers assigned to the Aila group (i.e. intention to treat), whether or not they used Aila during the trial.

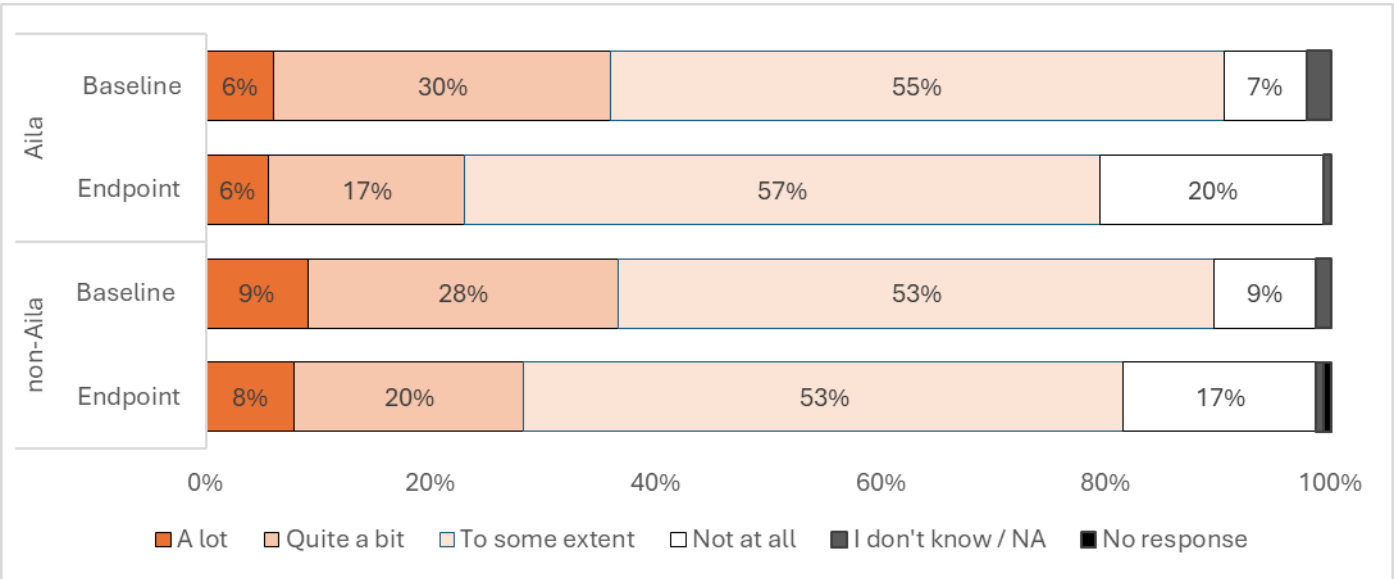
Figure 28: Probability of reduced, maintained and improved lesson planning workload for the Aila and non-Aila groups



Source: baseline and endpoint surveys, n = 300.

Based on the matched sample of teachers (n = 313) who responded to both the baseline and endpoint surveys, both Aila and non-Aila groups reported a modest reduction in lesson planning as helping to reduce stress in their work. The proportion of these teachers who reported that having too much lesson planning to do was a source of stress ‘a lot’ or ‘quite a bit’ reduced by from 36% to 23% (13 percentage points) for Aila-assigned teachers and from 37% to 28% (8 percentage points) for non-Aila assigned teachers. As this shift occurred in both groups, it is likely that this was primarily driven by other factors, such as the time of academic year. However, the improvement was slightly larger for the Aila group.

Figure 29: Extent to which lesson planning was a source of stress—baseline and endpoint surveys



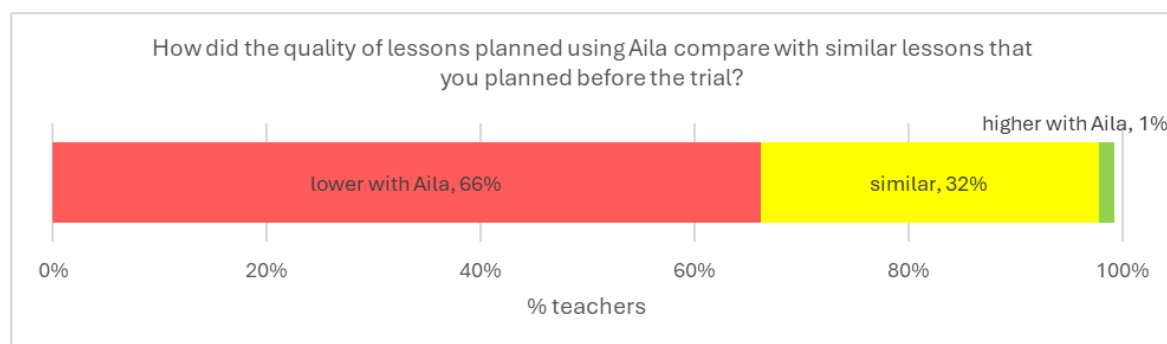
Source: baseline and endpoint surveys, n = 313.

Perceived impact on lesson quality, pupil engagement, and learning

Most Aila-assigned teachers who used Aila perceived that lessons planned with Aila were lower in quality than equivalent lessons planned before the trial (Figure 30). While teachers were asked to make this comparison based on the final resources used in the lesson (including their own edits), data from school visits suggests that teachers instinctively compared the raw Aila-generated resources (before their edits and additions) with their own fully planned lessons. Where teachers thought the resources as generated were not suitable for their class they amended them before use. (It should be noted that editing and adapting Aila resources prior to their use in the classroom is expected and encouraged by Oak National Academy.) Some teachers used only one or two elements of the Aila-generated resources (in conjunction with other resources) or integrated the Aila resources into a school lesson structure. Where teachers primarily used the Aila resources to teach with, amendments included changing starter quiz questions to reflect the actual prior learning of the class, adding images, and providing tailored examples.

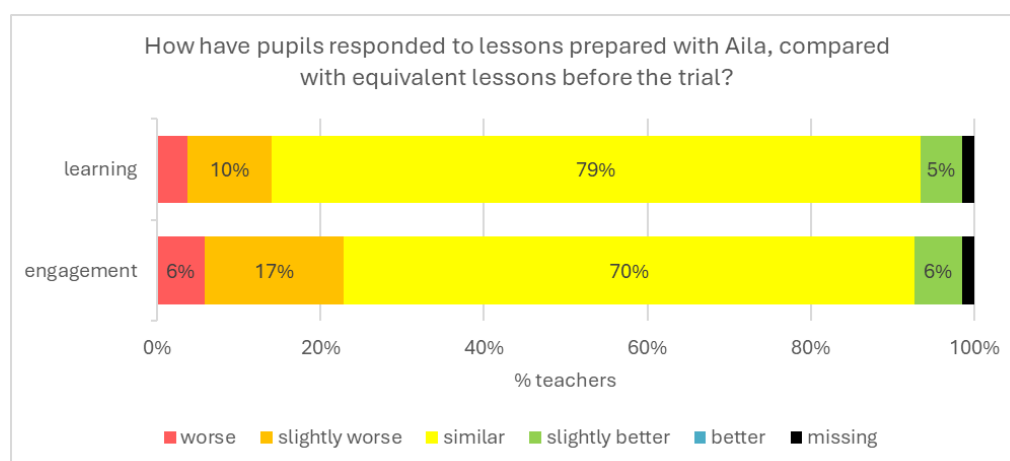
Pupil engagement and learning were felt to be similar to equivalent lessons before the trial (Figure 31). In the observed lessons ($n = 5$), teachers were in consensus that pupil engagement and learning were similar, whether or not Aila was used for planning.

Figure 30: Perceived quality of lessons (among Aila-assigned teachers who reported using Aila)



Source: endpoint survey, Aila-assigned teachers who reported using Aila, $n = 136$.

Figure 31: Perceived impact of Aila on pupil engagement and learning (among Aila-assigned teachers who reported using Aila)



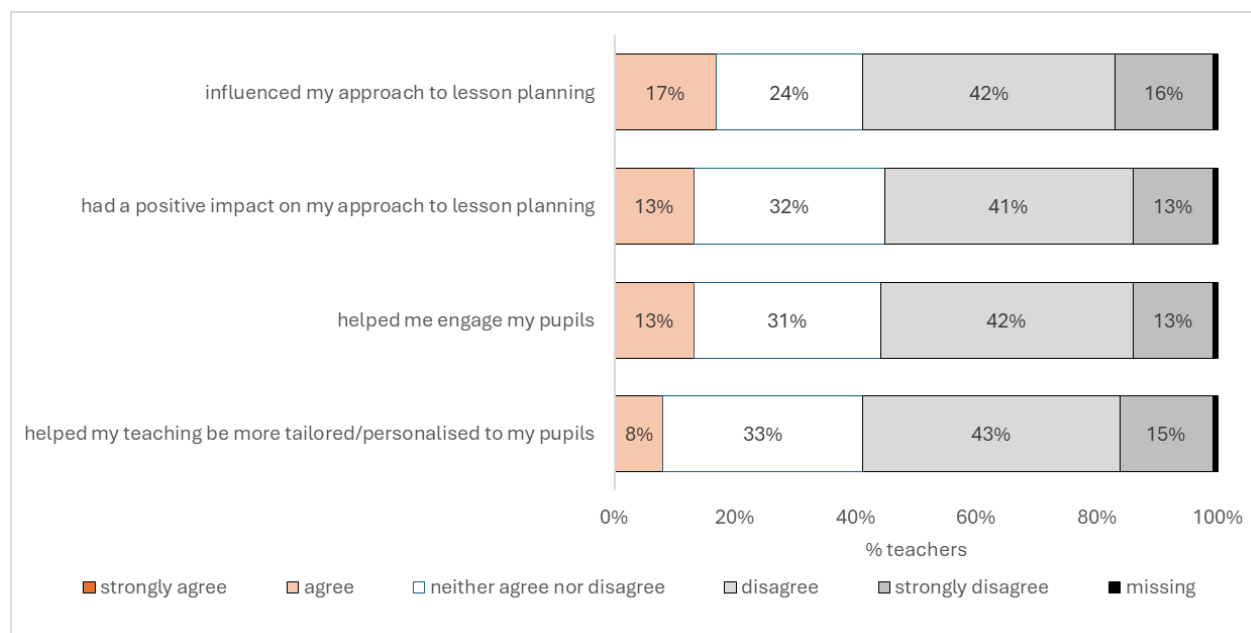
Source: endpoint survey, Aila-assigned teachers who reported using Aila, $n = 136$.

Perceived impact on lesson planning

Although Aila aims to model and scaffold an expert lesson planning process for teachers, a small majority of teachers felt that it had not had an impact on their planning and teaching. For example, over half disagreed that Aila had influenced their approach to lesson planning (58%), had had a positive impact on their lesson planning (54%), had helped them engage their

pupils (55%), or had helped their teaching be more personalised or tailored to their pupils (58%). Most other teachers neither agreed nor disagreed (24% to 32%), while less than a fifth agreed (8% to 17%). See Figure 32 for details.

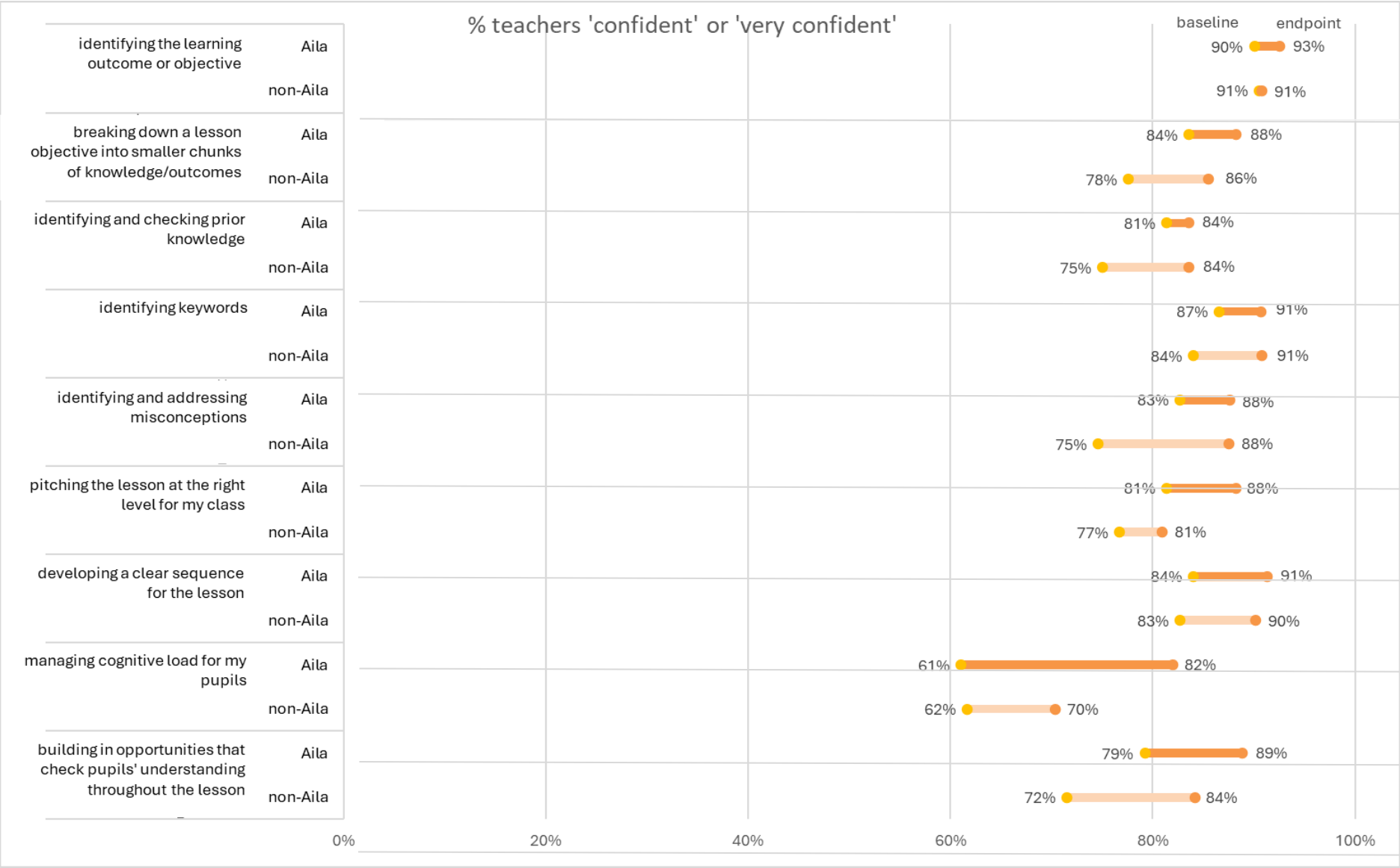
Figure 32: Teacher perceptions of the impacts of using Aila on lesson planning



Source: endpoint survey, Aila-assigned teachers who reported using Aila, n = 136.

This lack of a change may be related to how teachers were feeling at the start of the trial. Before the trial, most trial teachers reported they were already confident in the lesson planning processes modelled by Aila, consistent with the high levels of experience and broad confidence in lesson planning generally reported by teachers. Comparing the matched sample of teachers who completed both the baseline and endpoint surveys, teachers' confidence in all processes increased in both groups, though by differing amounts. Most processes showed similar increases for both the Aila and non-Aila groups, or minimal changes for both groups, suggesting no specific impact of Aila on these processes. However, two lesson planning processes did show a larger difference between groups. For 'managing cognitive load for my pupils', the Aila group had a larger gain than the non-Aila group (21 and 8 percentage points respectively). For 'identifying and addressing misconceptions' the non-Aila group had a slightly larger gain than the Aila group (13 and 5 percentage points respectively). See Figure 33 for details.

Figure 33: Perceived teacher impact on specific lesson planning processes



Source: baseline and endpoint surveys, matched sample, n = 313.

Perceived impact on financial costs

Almost all teachers (92%) reported that financial costs were similar when using Aila. Where teachers indicated reasons for higher costs (n = 2) this was due to extra printing. Similarly, some case study teachers commented on the volume of documents created by Aila, though they did not explicitly relate this to additional costs. Where teachers indicated reasons for lower costs (n = 4), this included savings from not subscribing to a paid GenAI tool, not purchasing other curriculum resources, or reduced printing.

Perceived impact and future intentions

Interviewed teachers who reflected positively on their experience of Aila were those who had seen positive impacts for themselves or their pupils. Two representative examples from the range of interviewed teachers are provided below.

Teacher C previously used another GenAI education tool to support planning, mainly to generate model texts for writing. During the trial, she used Aila where she wanted more reinforcement and consolidation activities to supplement existing plans, or to plan new lessons where she was less confident in subject knowledge for teaching.

She used Aila to plan a grammar lesson on punctuation for speech, which ‘my kids absolutely loved and actually really got them to understand the rules of speech. And ever since then, I’ve not had to remind children ... it broke it down into those step.’ This lesson ‘massively improved their writing’. Aila also suggested engaging scenarios for lessons, including a predict-observe-explain demonstration of balls of different sizes falling to the ground, which prompted ‘critical thinking’, and the jobs of different Anglo-Saxons in a village, which ‘really immersed the kids into that discussion’.

Aila ‘helped reduce my workload’, she reported, and influenced her approach to planning other lessons: ‘I’m definitely doing more exit quizzes towards the end of my lessons.’ She intends to continue using Aila, alongside another GenAI education tool which offers different functionality.

Teacher D previously used a general GenAI tool to support planning, for example, to generate retrieval and inference questions for specific reading texts. During the trial, she used Aila to support re-planning of a history unit and for spelling, punctuation, and grammar lessons.

In re-planning a unit on Mayan history, Aila was helpful in suggesting specific content and structure for each lesson, unlike the national curriculum, which does not provide detail on what to teach. Aila used an existing Oak National Academy curriculum unit as the ‘base’ material. Teacher D felt Aila helped her ‘break down that input as well, rather than giving the children all of the information at once’. She liked Aila’s emphasis on quizzes as ‘the going back and forth and the checking that they’ve understood keeps them engaged’.

For Teacher D, Aila ‘definitely reduced the amount of time that I’ve spent on the lessons that I’ve used it for, just because it’s given me that kind of basic structure ... [which] can take the longest, actually thinking, what do I want this lesson to look like?’ She intends to continue using Aila for re-planning lessons and planning ‘from scratch’.

Overall, 46% of Aila-assigned teachers expected to continue using Aila to some extent in the spring term (January to March 2026), however, among teachers who perceived positive impacts from Aila, the proportion who expected to use Aila in the following term was higher. This included teachers who perceived that:

- planning with Aila took less time (80% expected to use Aila next term, compared with 46% who reported similar time, and 30% who said it took more time);
- the quality of final lesson resources when planning with Aila was similar or higher (87% expected to use Aila) rather than lower (29%); and
- pupil engagement when planning with Aila was similar or better (57% expected to use Aila), rather than worse (19%).

Moderators

We hypothesised that teacher confidence in KS2 planning, subject knowledge for teaching, pedagogical confidence, and technological confidence would moderate teachers' use of Aila and the time they saved by using it. For example, we predicted that teachers with less teaching experience, lower confidence in planning, subject knowledge, and pedagogy, and lower prior use of GenAI might have higher use of Aila, while those with higher confidence in technology might have higher use of Aila and use it more effectively to save time. To explore this, the baseline survey asked about these areas of confidence. To contextualise the moderator analysis, we present an overview of the descriptive statistics from these questions.²²

Almost all teachers tended to strongly agree or agree that they have sufficient subject knowledge for their KS2 teaching (94%) and that they knew the basic theories and concepts of the KS2 subjects they teach (94%). Similarly, teachers broadly reported high pedagogical confidence, with more than three-quarters agreeing with each of a range of statements. See Table 32 for details.

Table 32: Baseline pedagogical confidence

Statement	% teachers 'strongly agree' or 'tend to agree'
I can adapt my teaching based upon what pupils currently understand or do not understand	95
I can adapt my teaching style to different pupils	92
I can assess pupil learning in multiple ways	91
I can use a wide range of teaching approaches in a classroom setting	91
I am familiar with common pupil understandings and misconceptions	91
I know how to select effective teaching approaches to guide pupil thinking and learning	90
I know how to develop appropriate tasks to promote pupils' complex thinking	76
I know how to develop exercises with which pupils can consolidate their knowledge	88
I know how to evaluate pupils' performance	92
I can explain essential content of my teaching subject in ways that pupils can understand	96

Source: baseline survey, n = 463.

Considering technological confidence, almost three-quarters (71%) said they had the technical skills needed to use technology, while less than half of teachers agreed that they frequently play around with technology (46%), kept up with important new technologies (42%), and knew a lot about different technologies (42%).

We conducted polychoric factor analysis across the items measuring confidence. The resulting factor structure, loadings, and proportion of variance are shown in Table 33. Only loadings with an absolute value larger than 0.3 are shown.

²² The descriptive statistics we present combine response categories for brevity. The factor analysis which underpins the moderator analysis uses the full individual response categories.

Table 33: Factor analysis for teacher confidence scales

Question Text	Scale	Factor 1	Factor 2	Factor 3
I have sufficient subject knowledge for my KS2 teaching	Agree/disagree	0.351		0.762
I know the basic theories and concepts of the KS2 subjects I teach	Agree/disagree	0.345		0.803
I can adapt my teaching based upon what pupils currently understand or do not understand	Agree/disagree	0.796		
I can adapt my teaching style to different pupils	Agree/disagree	0.746		
I can assess pupil learning in multiple ways	Agree/disagree	0.780		
I can use a wide range of teaching approaches in a classroom setting	Agree/disagree	0.767		
I am familiar with common pupil understandings and misconceptions	Agree/disagree	0.757		
I know how to select effective teaching approaches to guide student thinking and learning in my teaching subject	Agree/disagree	0.819		
I know how to develop appropriate tasks to promote students' complex thinking in my teaching subject	Agree/disagree	0.710		
I know how to develop exercises with which students can consolidate their knowledge of my teaching subject	Agree/disagree	0.748		
I know how to evaluate students' performance in my teaching subject	Agree/disagree	0.796		
I can explain essential content of my teaching subject in ways that students can understand	Agree/disagree	0.748		
I keep up with important new technologies	Agree/disagree		0.818	
I frequently play around with the technology	Agree/disagree		0.828	
I know about a lot of different technologies	Agree/disagree		0.877	
I have the technical skills I need to use technology	Agree/disagree		0.724	
	Proportion Var	39%	17%	10%
	Cumulative Var	39%	56%	66%

Source: baseline survey, n = 463.

Based on the factor loadings, we broadly interpret the factors as follows:

- Factor 1: pedagogical confidence;
- Factor 2: technological confidence; and
- Factor 3: subject knowledge for KS2 teaching.

We used these factors, plus a fourth model looking at teacher experience in terms of number of years teaching, to explore how teacher confidence at baseline is associated with the percentage of planning for which Aila is used. Table 34 shows the coefficient estimating the association between each factor and the percentage of planning with Aila within the statistical model (see IPE methods), along with *p*-values. None of these results reached the threshold for statistical significance, which means the results may have occurred by chance rather than reflecting an association.

Table 34: Association between teacher factors (baseline) and the percentage of Aila use

Factor	Factor interpretation	Group	Estimate of coefficient (95% CI)	p-value
Factor 1	Pedagogical confidence	Aila	0.15 (-0.14, 0.44)	0.316
Factor 2	Technological confidence	Aila	0.01 (-0.27, 0.28)	0.962
Factor 3	Subject knowledge for KS2 teaching	Aila	-0.23 (-0.51, 0.06)	0.121
Years of experience		Aila	-0.02 (-0.06, 0.02)	0.291

Source: diaries and baseline survey, n = 212.

The same measures were used to explore how teacher confidence at baseline moderated lesson planning workload. This analysis compares workload within and across groups for low factor scores (at the 10th percentile) and high factor scores (at the 90th percentile). Table 35 shows the high and low scores used for these comparisons.

Table 35: High and low scores for each factor/measure

Factor	Factor interpretation	Low score (10th percentile)	High score (90th percentile)
Factor 1	Pedagogical confidence	-1.039	1.368
Factor 2	Technological confidence	-1.290	1.200
Factor 3	Subject knowledge for KS2 teaching	-0.931	1.020
Years of experience		3	22

Source: diaries and baseline survey, n = 415.

Table 36 shows the fold change estimating the difference in lesson planning workload between intervention groups within low and high factor scores, and between low and high factor scores within each intervention group along with confidence intervals and *p*-values. When comparing teachers with low scores in each factor/measure across the Aila and non-Aila groups, we saw a significant coefficient of less than one for Factor 1 (pedagogical confidence), Factor 3 (subject knowledge for KS2 teaching) and years of experience. This means that teachers in the Aila group who reported lower pedagogical confidence, lower subject knowledge for teaching, or fewer years of teaching experience at baseline had significantly lower lesson planning workload, compared with similar teachers in the non-Aila group. *This suggests that Aila would reduce workload more for teachers with lower teaching experience and/or confidence.*

When comparing teachers with high scores in each factor across the Aila and non-Aila groups, we saw a significant coefficient of less than one for Factor 2 (technological confidence). This means that teachers in the Aila group with higher self-reported technological confidence at baseline had significantly lower lesson planning workload, compared with similar teachers in the non-Aila group. *This suggests that Aila would reduce workload more for teachers with higher technological confidence.*

Table 36: Association between teacher confidence (baseline) and lesson planning workload

Factor	Factor interpretation	Group	Estimate of fold change (95% CI)	p-value
Factor 1	Pedagogical confidence	Aila/non-Aila: low factor score	0.63 (0.48, 0.83)	0.000
		Aila/non-Aila: high factor score	0.95 (0.68, 1.31)	0.967
		Aila: high/low factor score	1.27 (0.92, 1.74)	0.208
		non-Aila: high/low factor score	0.85 (0.63, 1.15)	0.478
Factor 2	Technological confidence	Aila/non-Aila: low factor score	0.80 (0.59, 1.11)	0.276
		Aila/non-Aila: high factor score	0.71 (0.53, 0.96)	0.019
		Aila: high/low factor score	0.99 (0.72, 1.35)	1.000
		non-Aila: high/low factor score	1.12 (0.80, 1.56)	0.806
Factor 3	Subject knowledge for KS2 teaching	Aila/non-Aila: low factor score	0.68 (0.51, 0.91)	0.004
		Aila/non-Aila: high factor score	0.84 (0.62, 1.13)	0.420
		Aila: high/low factor score	1.19 (0.93, 1.53)	0.259
		non-Aila: high/low factor score	0.97 (0.70, 1.34)	0.994
Years of experience		Aila/non-Aila: low factor score	0.68 (0.51, 0.91)	0.003
		Aila/non-Aila: high factor score	0.87 (0.63, 1.20)	0.651
		Aila: high/low factor score	0.94 (0.69, 1.28)	0.947
		non-Aila: high/low factor score	0.74 (0.54, 1.01)	0.059

Source: diaries and baseline survey, n = 415.

Conclusion

Table 37: Key conclusions

Key conclusions	
1.	The lesson and resource preparation time spent by the Aila group was 76% of the time spent by the non-Aila group during the ten weeks of the trial. In this trial, teachers in the non-Aila group spent 3.3 hours per week preparing lessons and resources for their class, while teachers in the Aila group spent around 2.5 hours per week, saving 0.8 hours (or 48.6 minutes) per week on average.
2.	This reduction in teacher workload was evident without compromising the lesson resource quality: the lesson resource quality analysis showed that the quality of the lesson resources in the Aila group was no worse than that in the non-Aila group.
3.	Teachers in the Aila group with lower confidence in their subject knowledge, lower pedagogical knowledge, those new to the profession or with higher confidence in technology were more likely to have a lower workload than similar teachers in the non-Aila group.
4.	Teachers in the Aila group saved time even though Aila was used for an average of 7% of lessons. Notably, usage frequency and amount both declined over the ten weeks.
5.	Teachers had mixed views about the Aila tool: while they found it to be accessible and curriculum-aligned, many preferred alternative GenAI tools which provided more ease of use, and Aila, by design, seeks to assist teachers through the planning process, requiring more involvement. This may also reflect a disconnect between Aila’s intended purpose of supporting lesson planning from scratch and teachers’ planning practices as well as perceptions that its outputs often required adaptation before being classroom ready.

Impact evaluation and IPE integration

Evidence to support the logic model

The findings from this evaluation indicate that using Aila reduces teacher lesson planning time without compromising the quality of lesson resources, providing evidence in support of the hypothesised logic model, which has been annotated with key findings from the project. This pattern is consistent across primary and sensitivity analyses suggesting that the results are robust. There were similar baseline intended lesson planning levels between the two groups and teachers largely adhered to their allocated condition (although adherence was lower in the Aila group and the levels of Aila use fluctuated, and indeed declined, as described below).

The primary impact finding shows a significantly lower lesson planning time for teachers assigned to the Aila group than those in the non-Aila group during ten weeks of implementation suggesting that use of Aila had led to a reduction in workload. In real terms, teachers in the Aila group spent 24% less time on lesson planning meaning, on average, they saved 49 minutes per week on lesson planning compared to teachers in non-Aila group. This time saving result mirrors that of the previous ChatGPT trial (Roy et al., 2024). While direct comparisons of absolute time savings between the two trials are not appropriate given differences in key stages and timing of the trials, the relative reductions can be contrasted: in the ChatGPT trial, lesson preparation time was 69% of that of the comparison group, compared to 76% in the Aila trial, indicating a smaller relative reduction in this trial.

This effect is consistent across all sensitivity analyses, including exclusion of the first week (to allow for familiarisation) and exclusion of half-term weeks. The slightly stronger effect when excluding the first week suggests a learning or efficiency gain over time, indicating that teachers become more efficient at using Aila once they are familiar with the tool. However, although the majority of teachers in the Aila group tried Aila at least once (81% teachers), Aila was used infrequently during the trial period, accounting for only a small proportion of lessons (7% on average), and Aila use declined over time (53% of teachers reported their weekly planning included Aila in the first week, compared with 33% in week ten). This reduced use of Aila did not relate to reduced need for lesson planning as the weekly lesson planning time remained similar throughout the trial (with the exception of the half-term week). Overall, this indicates that the observed reduction in workload was achieved despite relatively low and selective use, rather than through sustained or comprehensive adoption. The pattern is

further supported by compliance analysis, which showed that teachers who used Aila more frequently experienced greater reduction in their lesson planning time.

As was the case in the ChatGPT trial, the main effect of reduced workload for the Aila group was despite low levels of usage of the tool. The ‘opportunity’ to use Aila was limited by the extent to which teachers needed to plan whole lessons, which is what the central function of Aila supports teachers to do. ‘Capability’ appeared to be less of an issue but ‘motivation’ waned as the trial period progressed. Teachers had mixed views overall about the ease of using different elements of Aila, and about half intended to continue to use Aila after their participation in the trial was over. The IPE findings indicated that when teachers did use Aila they mainly used the function for planning whole lessons—rather than the resource generation tool, which can produce stand-alone glossaries, comprehension tasks, and quizzes—even if they then only used some of the resources Aila generated, supplemented by information from other sources including their own knowledge and experience.

The limited use of Aila was further illustrated by the reasons given by teachers who stopped using it during the trial: (i) limited need whereby teachers did not need to plan from scratch or they needed question banks, which are not explicitly supported by Aila; (ii) negative perceptions of experiences with Aila such as describing it as ‘slow’, ‘inflexible’, and ‘cumbersome to use’, with outputs often perceived as formulaic, lacking detail and engagement, and requiring substantial time to adapt before use; and (iii) preference for a different GenAI tool as it better fitted their planning needs. Together, this reflects gaps in elements of the ‘technical taxonomies’ (see Behavioural Insights, 2025) in relation to use of AI, specifically in terms of:

- (i) demonstration of low ‘relative advantage’ compared to existing planning approaches;
- (ii) limited ‘compatibility’ with a person’s values and norms—we may assume that participants were in favour of AI as they agreed to be part of the project but this was not measured directly; and
- (iii) a lack of ‘observability’ wherein benefits were perhaps not seen clearly (though this may change in some ways with the publication of the impact results in this report).

Teachers who perceived they saved time were, naturally, more likely to report they would continue to use Aila in the future. Positive elements of Aila included being able to use the tool as much or as little as they liked and while less than one in five teachers agreed that it had influenced their teaching and planning, teachers did appreciate that it supported them to plan lessons in new topics and suggested new ideas for their teaching. However, the actual Aila usage was lower in the spring term (after the trial ended) compared to initial intentions to continue using Aila after the end of the trial.

The findings show that lessons produced in the Aila group were not significantly worse than those in the non-Aila group, meeting the predefined non-inferiority margin (though it should be noted that this was also an intention-to-treat analysis, that is, the resources analysed were the ‘end product’ used by the teachers in the classroom including any adaptations of Aila outputs by teachers, and which may have also drawn on information from other sources—only 71% of which were noted as being created using Aila). The lesson resource quality analysis indicates that teachers can achieve workload reductions without compromising the quality of lesson materials, which is a key consideration for practice. However, when asked, most teachers who used Aila perceived the quality of its lesson resources to be lower than their previous planning approaches, citing concerns such as inflexibility, lack of detail, and overly formulaic outputs. These findings suggest a contrast between perceived and evaluated lesson resource quality. The Aila-generated resources were seen by teachers as a starting point—which fits with the intention of Oak National Academy—however, teachers’ perceptions also indicated that the Aila-generated resources may not always align with their expectations of flexibility, relevance, or level of detail, which meant teachers had to make greater adaptations to these resources before they were used in classrooms than anticipated, for example, to make them suitable for their pupils and curriculum context.

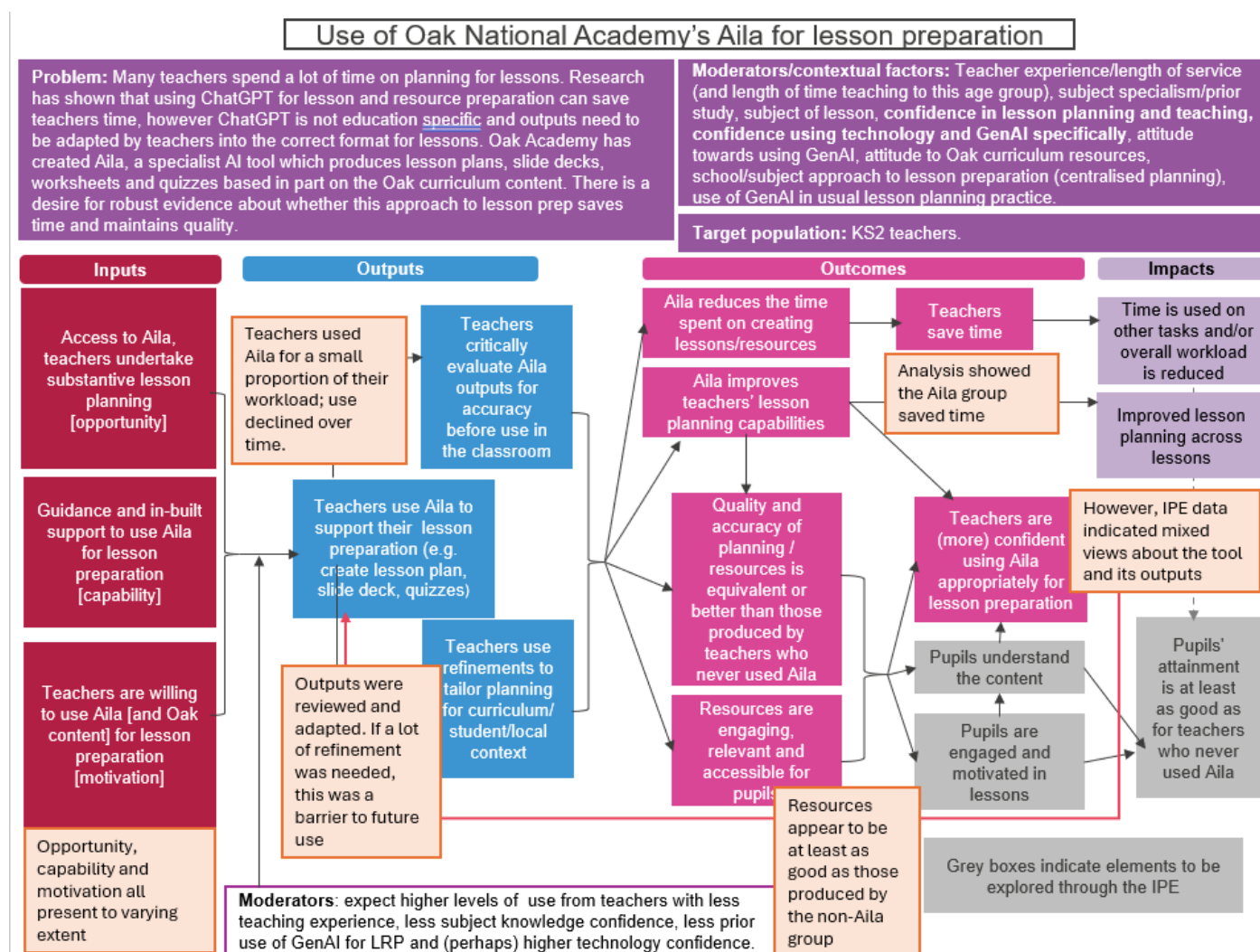
Moderators and differential impact

Overall, the subgroup analyses provide limited evidence of differential effects of Aila on workload based on either prior GenAI use or the proportion of lessons planned from scratch. Although some individual subgroup models suggested larger effects (for example, for teachers with no prior GenAI use), these findings were based on small samples and should therefore be interpreted with caution. Overall, the results suggest that workload reductions were broadly similar across teachers rather than driven by specific subgroups.

Teachers in the Aila group with lower confidence in their subject knowledge, lower pedagogical knowledge, and those new to the profession were more likely to report a lower workload than similar teachers in the non-Aila group. This suggests that Aila may be particularly beneficial at saving time for teachers at early stages of their career or who have lower confidence in their subject knowledge and pedagogy. However, this does not necessarily imply that the quality of the lesson delivered is as good as if it had been planned without Aila. Similarly, teachers in the Aila group with higher confidence in using technology were more likely to report lower workload than similar teachers in the non-Aila group, a finding in line with the ChatGPT trial reported in 2024. This finding suggests that Aila would reduce workload more for teachers with higher technological confidence.

Figure 34: Updated logic model

Orange boxes indicate where evidence supports/does not support the hypothesised logic model.



Interpretation

The results of this trial provide additional evidence to the growing literature around how GenAI tools can support teachers to reduce their workload. Even with relatively low levels of Aila use, the Aila-assigned teachers saved a considerable amount of time per week compared with the non-Aila group. However, although the impact analysis showed teachers in the Aila group spent less time on planning lessons, this does not appear to have translated into how teachers feel about their workload as most teachers that used Aila felt the time to plan lessons was about the same or took longer. It is important to consider the comparisons being made here: the impact analysis is making a contrast between two randomly allocated

groups whereas the IPE is based on self-reported perceptions of how the time felt compared to their own 'non-Aila' previous planning approaches. This also extended to perceptions about the content and use of the tool: the IPE indicated that while some teachers found Aila beneficial, most participating teachers did not find Aila quicker or easier to use than other sources, including other forms of GenAI.

The literature provides perspective on how teachers perceive and rate sources of information and has previously highlighted a disconnect between objective quality of resources, their relative popularity, and teachers' preferences. Previous research has shown that teachers can view teaching as 'an art not a science', suggesting that part of the process of planning for a lesson is their craft and has elements of creativity in terms of drawing together different ideas and sources into a single lesson (Bugler, 2017). Other research has indicated that teachers factor in ease of access (among other factors) when rating the quality of resources more highly than content, at least during the initial stages of selection (Polikoff and Dean, 2019). Together, this may provide some context for why the impact and IPE results appear to be in conflict.

Where Aila was used, teachers tended to use the function for planning a whole lesson (even though there is a function to plan three types of resources without planning a whole lesson), however, the requirement for an entirely new lesson was relatively infrequent so teachers appeared to be going through all the stages and requirements for planning a whole lesson and then discarding parts they did not need. This may also explain some of the negative attitudes towards how long the process was perceived to take compared to 'usual' planning where they are less likely to complete superfluous steps.

Findings from the light-touch scoping activities also provide context for interpreting the findings from this trial. Teachers reported using a combination of individual and collaborative planning approaches, with high levels of autonomy in selecting resources. This indicates that lesson planning practices are diverse and flexible. In addition, a substantial proportion of teachers were already using some form of GenAI tools, although the use had not yet reached saturation. This suggests that Aila needed to 'slot in' amongst established and evolving lesson planning practices, which could help explain some of the reasons behind why it was typically used to supplement rather than replace other tools, and why overall usage declined over time. Perhaps, in the most part, teachers value tools which can be used flexibly to enhance the lesson planning process rather than to plan an entire lesson from scratch.

Other research suggests that the use of GenAI for lesson planning (such as idea generation), not simply task execution (for example, quiz generation), are more likely to see productivity gains (Keppler, Sinchaisri and Snyder, 2024).

Previous research (for example, Roy et al., 2024) has shown that some users of GenAI interact with it only superficially, similar to the way they would use a search engine, rather than utilising the full capabilities through refinements and tailored prompts. Aila-assigned teachers reported making relatively high levels of refinements, both within Aila (using the prompts, and as scaffolded by Aila) and after downloading the Aila resources. While this appears to be an improvement on previous GenAI use by moving to more sophisticated interactions with AI, only around a third of teachers found this process 'easy'.

Aila is freely available and therefore the choice conditions (Aila and non-Aila) would be possible to replicate in day-to-day practice, or indeed in another trial.

Overall, this report provides evidence to support school senior leaders' and teachers' decisions around workload and use of GenAI. While the use of AI can reduce workload and maintain quality of outputs, the IPE illustrates that it is important that teachers have agency over which sources and resources they can draw upon and evidence about the relative effectiveness. Participants in this trial were not aware of the headline finding of time saving, nor indeed by how much: it remains to be seen whether having this information may have improved their perceptions of using the tool and whether it would have encouraged higher levels of use.

Limitations and lessons learned

The estimated effect reflects the added value of Aila relative to existing teacher practice—where GenAI is already embedded in lesson planning process—rather than compared to no use of GenAI. While this limits causal isolation of Aila's standalone effects (in the absence of any GenAI tool), it is consistent with a 'real-world' Teacher Choices design and therefore these findings have practical relevance.

There are considerations relating to generalisability. The sample excluded teachers who were already frequent users of Aila or those who wished to continue using new resources from Oak National Academy meaning that the evaluation does not capture the experiences of more established users. While the IPE provides insight into teachers with mixed or less favourable experiences with Aila, it may under-represent a group of practitioners who use the tool extensively and effectively. As such, the findings are most applicable to teachers new to Aila or using it at low levels and may not provide useful conclusions about its impact among more established users of Oak National Academy resources, including Aila. Previous research conducted by Oak National Academy reported high levels of positive perceptions of the tool and quality of the outputs. The trial also had a lower proportion of teachers who had never used GenAI for lesson planning compared with the teacher population (24% in the trial, compared with 35% in a national survey from a similar timepoint). This means it under-represents teachers who are new to using GenAI for planning who could benefit differently from Aila's scaffolded approach. Qualitative samples for interviews and case study visits were small and were selected purposively. For example, school visits were undertaken to teachers who had indicated using Aila in the first three weeks of the trial and teachers were asked to plan the observed lesson with some use of Aila.

The primary outcome measure of lesson planning workload was self-reported, meaning that responses may have been susceptible to bias. In particular, it is possible that teachers' responses were influenced by Oak National Academy's communications highlighting potential time-saving benefits. These emails were sent out as standard to everyone using Oak National Academy's resources, not as part of this evaluation. While this may have influenced how teachers reported their lesson planning workload, several factors mitigate this risk. Diary questions were neutrally worded and administered consistently across both randomised groups. Furthermore, respondents were asked about the actual time they had spent on lesson planning in the previous week and were encouraged to respond honestly to diary questions (see Appendix J for exact wording). There is no direct evidence that responses were biased but we can't rule it out.

The secondary outcome measure of lesson resource quality is subject to measurement limitations whereby it was assessed using a structured rubric by an independent panel of KS2 teachers, where the inter-rater agreement was moderate. This indicates subjectivity in judgements of lesson resource quality. Besides, the ICC for the Aila group was lower than for the Non-Aila group, meaning reviewers were more likely to disagree on lesson quality for Aila group lessons or that the underlying quality of the Aila group lesson plans was similar within the reviewed set—or a combination of both. While this variability was accounted for in the statistical analysis, it reduces confidence in detecting small differences between groups. Our recommendation for future, similar, research would be to return to a ranking approach (see Roy et al., 2024) while ensuring enough resources are collected to sample from. A comparative judgement approach could also be considered if practical and feasible.

There are also limitations in the availability and completeness of usage data and the inability to match this to the IPE data. It is likely that the Aila platform data extracted from Oak National Academy did not capture all instances of use as the usage data was lower than the self-reported levels of use in the diary. While this is, at least in part, likely to be due to teachers using multiple or different accounts to access Aila (despite the research team's efforts to collect all logins from participants), there may also be an element of social desirability bias when participants were responding to the question in the diary, which asked, 'Did you do the thing we asked you to do as part of the trial.' Although this does not affect the primary outcome result, which was conducted on an intention-to-treat basis, this has implications for interpretation of impact as it appears that estimates of Aila usage may be affected.

There was some overlap in the use of Oak National Academy resources across both randomised groups, even though all trial participants were asked to refrain from using resources from the main Oak National Academy website (unless they had downloaded them before the trial). Reported use was generally limited and does not directly overlap with Aila's functionality in a way that would bias the comparison. Indeed, this reflects real-world practice where teachers draw on multiple sources in their lesson planning process wherein Aila is one such tool.

Although it would be interesting to be able to link 'time saved' to lesson quality, to minimise the data collection burden on teachers (given how underpowered this analysis would be) we did not collect the exact planning date nor time taken for individual lesson resources, which means it is not possible to make this connection. Time saved calculations from the impact analysis are based on group averages and therefore not linked to individuals nor their submitted resources.

As with the ChatGPT trial, this trial was an opportunity to understand teachers' decisions and patterns of use when 'encouraged' to use a particular tool. Fidelity and adherence are, therefore, conceptualised very differently to a typical programme evaluation and we see immense value—for Teacher Choices trials—in understanding impact given real-world choices about use (including when not to use) rather than insisting that participants must use their assigned tool (and, for example, structuring the incentives accordingly), which would be a different research question.

Future research and publications

This is the second of two Teacher Choices trials exploring the impact of using GenAI tools for lesson planning and resource preparation, both of which have found a positive effect on workload despite proportionally moderate or low levels of use of the tool. These trials were in different phases of education and across different subjects suggesting that there may be some applicability beyond the types of teachers and lessons involved in each project. This is still a relatively new area of research and the evidence base can still be strengthened, including by exploring other areas of teacher workload that AI can be used for, such as marking, feedback, and administrative tasks. As teachers varied in their extent of use and perceived impact of Aila, future research could profitably explore which teachers GenAI tools benefit the most, for example, early career teachers. Further research may also be useful to understand the influence of using AI on teachers' and pupils' understanding of topics and effects on pupil learning when teachers use AI frequently to plan their lessons. Future research could usefully examine how teachers modify and adapt AI-generated outputs, including a closer look at the adaptation process and lesson planning workflow.

References

- Behavioural Insights (2025) 'Adopt: AI and Human Behaviour': <https://www.bi.team/wp-content/uploads/2025/09/BIT-AI-ADOPT-2025.pdf>
- Benton, L. (2025) 'How is Aila Impacting Teacher Lesson Planning Practices, Workload and Expertise? Early Insights from the Classroom', Oak National Academy: <https://www.thenational.academy/blog/how-is-aila-impacting-teacher-lesson-planning-practices-workload-and-expertise-early-insights>
- Bugler, D. (2017) 'How Teachers Judge the Quality of Instructional Materials': https://assets.publishing.service.gov.uk/media/60795936d3bf7f400b462d74/Early-Career_Framework_April_2021.pdf
- Buluş, M., Dong, N., Kelcey, B. and Spybrook, J. (2021) 'PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments': <https://doi.org/10.32614/CRAN.package.PowerUpR>.
- Cicchetti, D. V. (1994) 'Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.' *Psychological Assessment*, 6(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Clark, H. B. (2025) 'Your Top 5 Questions About Aila, Our AI Lesson Assistant, Answered', Oak National Academy (9 April 2025): <https://www.thenational.academy/blog/your-top-5-questions-about-aila-our-ai-lesson-assistant-answered>
- DfE (2014) 'National Curriculum', London: Department for Education. <https://www.gov.uk/government/collections/national-curriculum>
- DfE (2025a) 'Generative Artificial Intelligence (AI) in Education', London: Department for Education. <https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education/generative-artificial-intelligence-ai-in-education>
- DfE (2025b) 'Technology in Schools Survey: 2024 to 2025'. London: Department for Education. https://assets.publishing.service.gov.uk/media/692834a6ce50d215cae9610e/Technology_in_schools_survey_2024_to_2025_research_report.pdf.
- DfE and OIT (2024) 'Generative AI in Education: Educator and Expert Views': https://assets.publishing.service.gov.uk/media/65b8cd41b5cb6e000d8bb74e/DfE_GenAI_in_education_-_Educator_and_expert_views_report.pdf
- DSIT, DfE, Kyle, P. and Morgan, S. (2024) 'Teachers to Get More Trustworthy AI Tech, Helping Them Mark Homework and Save Time': <https://www.gov.uk/government/news/teachers-to-get-more-trustworthy-ai-tech-as-generative-tools-learn-from-new-bank-of-lesson-plans-and-curriculums-helping-them-mark-homework-and-save>
- EEF (2022) 'Statistical Analysis Guidance for EEF Evaluations', London: Education Endowment Foundation. <https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1699621596>
- Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A.-W. and Michie, S. (2014) 'Better Reporting of Interventions: Template for Intervention Description and Replication (TIDieR) Checklist and Guide', *BMJ*, 348 (March 07, 3), pp. g1687–g1687. <https://doi.org/10.1136/bmj.g1687>
- IFF Research and UCL Institute of Education (2025) 'Working Lives of Teachers and Leaders: Wave 4, Summary report': https://assets.publishing.service.gov.uk/media/693bdc426a12691d48491e9b/Working_lives_of_teachers_and_leaders_wave_4_summary_report.pdf
- Keppler, S., Sinchaisri, W. P. and Snyder, C. (2024) 'The Teacher's Knapsack: Generative AI and Teacher Productivity, SSRN': <https://doi.org/10.2139/ssrn.4924786>
- Laird, N. M. and Ware, J. H. (1982) 'Random-Effects Models for Longitudinal Data', *Biometrics*, 38 (4), pp. 963–974. <https://doi.org/10.2307/2529876>.
- Mayer, R. (ed.) (2005) *The Cambridge Handbook of Multimedia Learning (1st edn.)*, Cambridge University Press. <https://doi.org/10.1017/CBO9780511816819>

- McLean, D. and Worth, J. (2025) 'Teacher Labour Market in England: Annual Report 2025':
https://www.nfer.ac.uk/media/afsn0rmb/teacher_labour_market_in_england_annual_report_2025.pdf
- Michie, S., van Stralen, M. M. and West, R. (2011) 'The Behaviour Change Wheel: A New Method for Characterising and Designing Behaviour Change Interventions', *Implementation Science*, 6 (42). <https://doi.org/10.1186/1748-5908-6-42>
- Oak National Academy (2025) 'AI Experiments: Build Tailor-Made Lessons and Teaching Materials with AI':
<https://labs.thenational.academy/>
- OECD (2018) 'TALIS 2018 Database': <https://www.oecd.org/en/data/datasets/talis-2018-database.html>
- Polikoff, M. and Dean, J. (2019) 'The Supplemental Curriculum Bazaar: Is What's Online Any Good?', Washington, DC: The Thomas B. Fordham Institute. <https://fordhaminstitute.org/national/research/supplemental-curriculum-bazaar>
- Rosenshine, Barak (2012) 'Principles of Instruction: Research-Based Strategies that All Teachers Should Know', *American Educator*, 36 (1), pp. 12–20.
- Roy, P., Poet, H., Staunton, R., Aston, K. and Thomas, D. (2024) 'ChatGPT in Lesson Preparation: a Teacher Choices Trial':
https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/chatgpt_in_lesson_planning_-_evaluation_report.pdf?v=1734340104
- Schafer, J. L. (1999) 'Multiple Imputation: A Primer', *Statistical Methods in Medical Research*, 8 (1), pp. 3–15.
<https://doi.org/10.1177/096228029900800102>
- Schmid, M., Brianza, E. and Petko, D. (2020) 'Developing a Short Assessment Instrument for Technological Pedagogical Content Knowledge (TPACK.Xs) and Comparing the Factor Structure of an Integrative and a Transformative Model', *Computers and Education*, 157, p. 103967. <https://doi.org/10.1016/j.compedu.2020.103967>
- Schmidt, A. F. and Finan, C. (2018) 'Linear Regression and the Normality Assumption', *Clinical Epidemiology*, 98, pp. 146–151. <https://doi.org/10.1016/j.jclinepi.2017.12.006>
- Shearing, H. (2025) 'Teachers Can Use AI to Save Time on Marking, New Guidance Says', BBC (10 June 2025):
<https://www.bbc.com/news/articles/c1kvj7dkp0o>
- Strom, R. (2025) 'Introducing Aila, Our AI-Powered Lesson Assistant', Oak National Academy:
<https://www.thenational.academy/blog/introducing-aila-for-ai-lesson-planning>
- Sweller, J. (1988) 'Cognitive Load During Problem Solving: Effects on Learning', *Cognitive Science*, 12 (2), pp. 257–285.
https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J. (2005) 'Implications of Cognitive Load Theory for Multimedia Learning', in R. Mayer (ed.), *The Cambridge Handbook of Multimedia Learning (1st edn)*, Cambridge University Press (pp. 19–30).
<https://doi.org/10.1017/CBO9780511816819.003>
- Teacher Tapp (2025a) 'AI Spy: Tracking the AI Tools Teachers Use': <https://teachertapp.com/app/uploads/2025/01/AI-Spy-Tracking-the-AI-Tools-Teachers-Use.pdf>
- Teacher Tapp (2025b) 'AIs Ups and Downs, Mobile Phones Plus Flexible Wish List...': <https://teachertapp.com/articles/ais-ups-and-downs>
- Teacher Tapp (2025c) 'The Rise of "AI Enabled" Teachers, Stressed Teachers, and Our Tapp Behaviour Tracker':
<https://teachertapp.com/articles/the-rise-of-ai-enabled-teachers-stressed-teachers-and-our-tapp-behaviour-tracker/>
- Walker, M., Worth, J. and Van den Brande, J. (National Foundation for Education Research) (2019) 'Teacher Workload Survey 2019: Technical Report', Department for Education:
https://assets.publishing.service.gov.uk/media/5e12fcc9e5274a0fa4dc2894/teacher_workload_survey_2019_technical_report__amended.pdf#page=16 (Accessed: 18 September 2025).

Appendix A: Security classification of trial findings

PRIMARY OUTCOME: Weekly hours spent in lesson and resource preparation, from researcher-designed weekly teacher diary

Rating	Criteria for rating			Initial score	Adjust	Final score
	Design	MDES	Attrition			
5 	Randomised design	≤ 0.2	0-10%			
4 	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%	4		4
3 	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%		 Adjustment for threats to internal validity [0]	
2 	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%			
1 	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%			
0 	No comparator	≥ 0.6	$>50\%$			

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Low	Randomised design with baseline adjustment and broadly balanced groups.
Threat 2: Concurrent Interventions	Low	Other GenAI use was permitted and common in both groups. This is appropriate for Teacher Choices, and the test is Aila versus usual practice including other GenAI, not Aila versus no GenAI.
Threat 3: Experimental effects	Low/Moderate	Participating teachers were aware of their allocation. Some evidence of contamination with 14% of teachers in the non-Aila group having accessed Aila but reported average Aila use among this cohort was low (0.6% of lessons). As Aila is freely available, this was anticipated by the evaluators and accounted for in their two-sided CACE analyses, the results of which support the primary analysis findings. Oak Academy newsletter communications about time-savings may have reached participants and influenced their perceptions, although direct evidence is not available.
Threat 4: Implementation fidelity	Moderate	81% of teachers in the Aila group reported using the tool for an average of 7% of lessons, and use declined over time. Teachers were free to use Aila as much as they wished and given the tool's intended purpose and the lesson planning opportunities for KS2 generalist teachers, this level of adoption is not entirely unexpected.
Threat 5: Missing Data	Low	Overall teacher attrition was 10.3%. A teacher's status as trial lead was noted as a predictor of missingness. Estimates from missing data and sensitivity analyses were reassuringly comparable to the primary analysis estimate.
Threat 6: Measurement of Outcomes	Moderate	The primary outcome is self-reported weekly workload in an unblinded trial, but the measure is based on the validated 2018 TALIS. Based on a non-normal distribution of the primary outcome data, a log transformation (pre-specified contingency) was applied, and which resulted in a broadly acceptable distribution.
Threat 7: Selective reporting	Low	Material study-plan deviations are reported and limitations are acknowledged.

- **Initial padlock score:** 4 padlocks- Well-conducted two-armed Teacher Choices randomised controlled trial. The MDES at randomisation stage was 0.22 and there was 10.3% teacher-level attrition from the trial.
- **Reason for adjustment for threats to validity:** [0] Padlocks – There are 3 potential moderate risk ratings. Experimental effects represent a low to moderate threat given evidence of contamination with 14% of teachers in the non-Aila group reporting use of the tool for 0.6% of lessons, and this may have reduced the observe difference between groups. Implementation fidelity is assessed as a low to moderate risk because adoption was partial and use of Aila declined overtime, potentially diluting the primary impact estimate. Finally, measurement of outcomes is assessed as a moderate risk given the use of a self-report measure in an unblinded trial. On the evidence reviewed, no further downward adjustment.
- **Final padlocks score:** 4 padlocks.

Appendix B: Tables and Charts for impact evaluation results

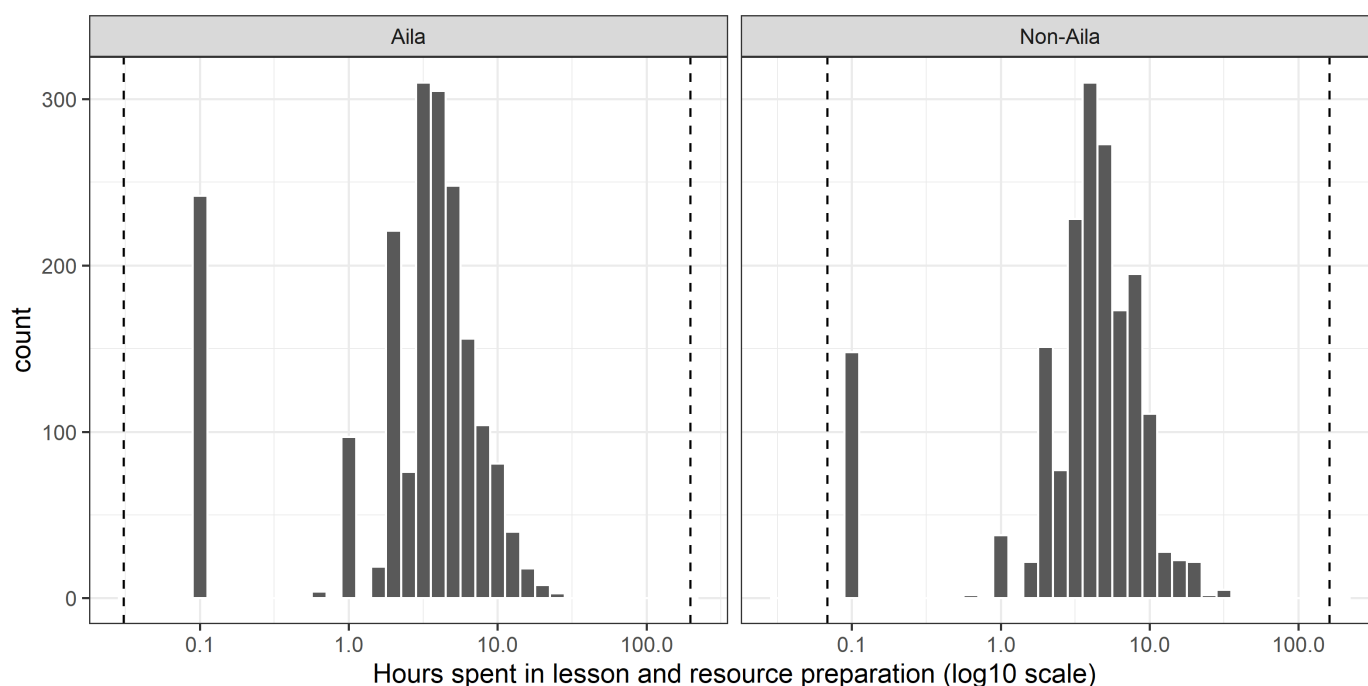
B.1: Presence of ‘extreme’ values

For the RQ1 outcome, weekly hours spent in lesson and resource preparation, the frequency of values outside of the range defined by the mean ± 3.29 *standard deviation (Tabachnick and Fidell, 2013) is reported for both randomised groups. Due to the skewed distribution of the data and scale that the data were on for analysis, the mean and standard deviation were calculated on the $\log_{10}$ transformed data. The mean and limits were then back transformed to the hours scale.

Table 38: ‘Extreme’ value limits and frequency of values outside the range

Group	Mean (Lower limit, Upper limit) – $\log_{10}$ scale	Mean (Lower limit, Upper limit) – hours scale	Number of data points within limits	Number of data points below lower limit	Number of data points above upper limit
Aila	0.390 (-1.51, 2.29)	2.46 (0.031, 196.6)	1933	0	0
Non-Aila	0.522 (-1.16, 2.21)	3.33 (0.068, 161.9)	1808	0	0

Figure 35: Weekly hours spent in lesson and resource preparation with ‘extreme’ value limits shown as vertical dashed lines



B.2: Outcome distributions

Baseline

Figure 36: Histogram of raw baseline data—expected weekly hours of lesson and resource preparation

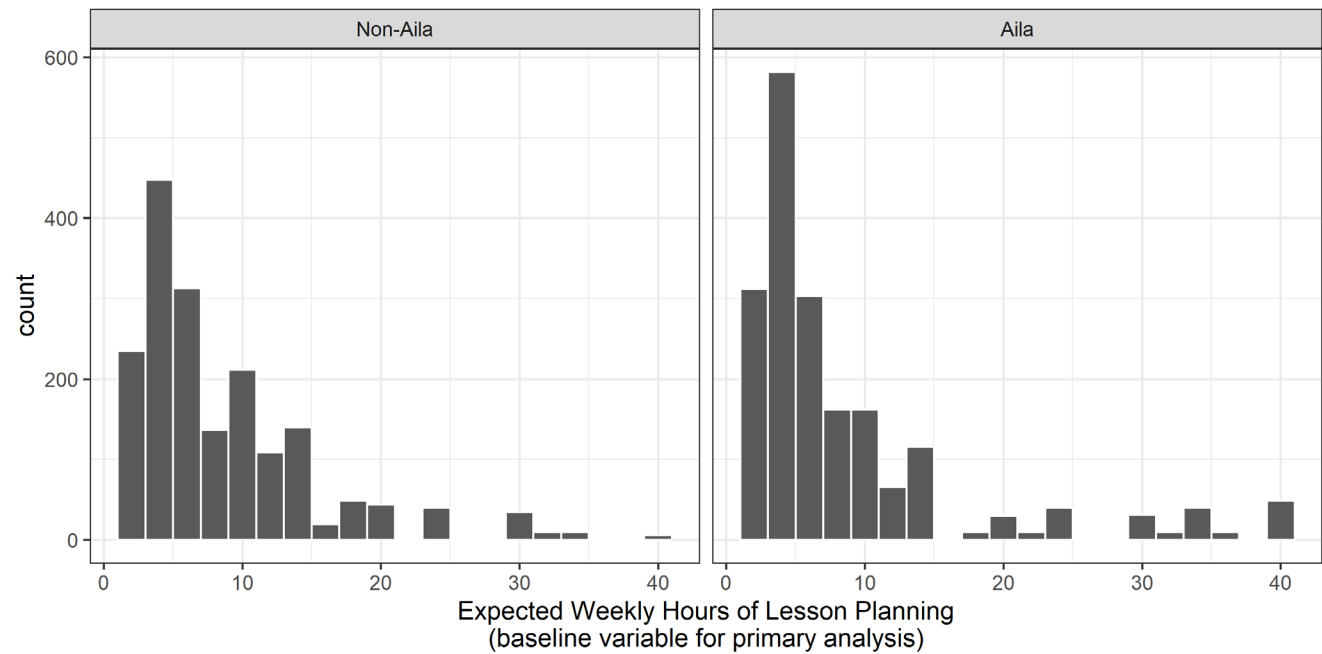
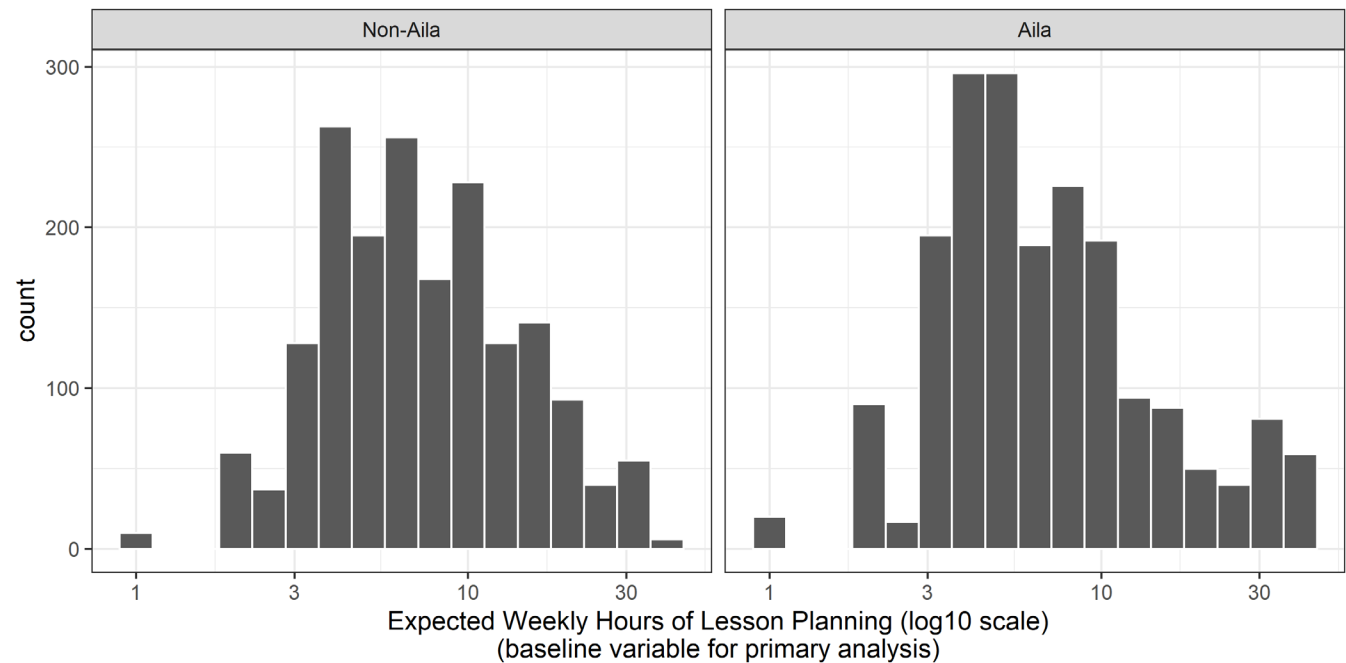
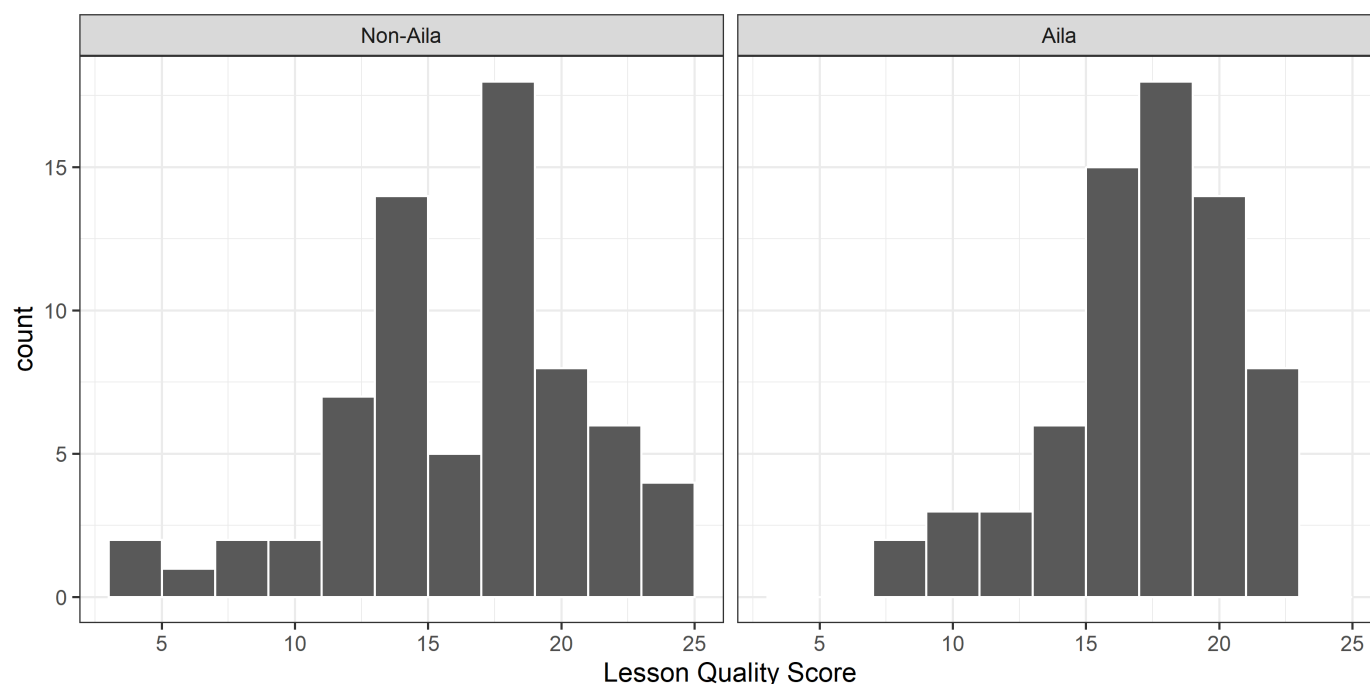


Figure 37: Histogram of $\log_{10}$ transformed baseline data—expected weekly hours of lesson and resource preparation



Lesson quality

Figure 38: Distribution of score assigned to lesson resources, split by randomisation group



B.3: Effect size calculation

Table 39 below shows several model parameters (difference in adjusted means, pooled variance) and several parameters calculated from raw data (difference in unadjusted means, variance of outcome). As described in the methods, the effect size is the adjusted difference in means on the $\log_{10}$ scale, divided by the square root of the total variance from the null model.

Table 39: Effect size calculation

			Aila group		Non-Aila group		
Outcome	Difference (Aila – Non-Aila) in unadjusted means	Difference (Aila – Non-Aila) in adjusted means	n (missing)	Variance of outcome	n (missing)	Variance of outcome	Pooled variance (square root of the total variance from the null model)
Weekly hours spent in lesson and resource preparation	-0.874	-0.816	214 (18)	10.9	202 (30)	13.8	3.54
Weekly hours spent in lesson and resource preparation – $\log_{10}$ scale, as analysed	-0.132	-0.121	214 (18)	0.335	202 (30)	0.263	0.550

B.4: ICCs

Table 40: ICCs and other model parameters

Research question	Number of schools	Number of teachers	Number of data points (weeks)	Between school variance	Between teacher variance	Within teacher variance	School ICC	Teacher ICC	R2 overall	R2 between	R2 within
RQ1	106	416	3741	0.00643	0.0715	0.207	0.0226	0.251	0.0634	0.160	0.0368
RQ3	106	413	3569	0.00730	0.0765	0.210	0.0248	0.260	0.0683	0.169	0.0386
RQ4	106	416	3337	0.00619	0.0789	0.111	0.0315	0.402	0.0873	0.154	0.0497
RQ5a – interaction	106	416	3741	0.00668	0.0722	0.207	0.0234	0.253	0.0643	0.220	0.0388
RQ5b – interaction	106	416	3741	0.00345	0.0744	0.207	0.0121	0.262	0.0807	0.427	0.0475
Missing data - sensitivity analysis	106	416	3741	0.00629	0.0712	0.207	0.0221	0.251	0.0634	0.160	0.0368
Missing data - complete case analysis	100	315	3150	0.01062	0.0743	0.209	0.0362	0.253	0.0638	0.207	0.0288

Table 41: ICCs and other model parameters for RQ2

Research question	Number of reviewers	Number of lessons	Between reviewer variance	Within reviewer variance	Reviewer ICC	R2 overall	R2 between	R2 within
RQ2	6	42	0.00	16.4	0.00	0.0156	0.157	0.0168

Further appendices

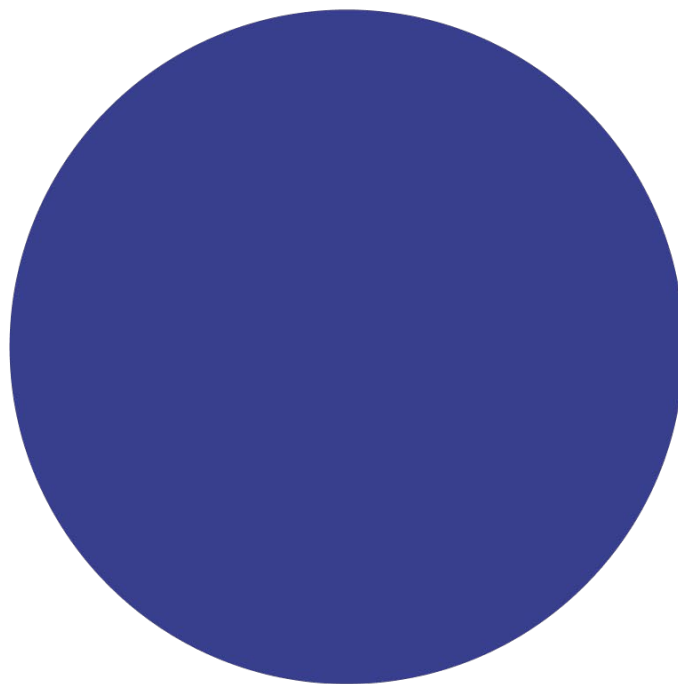
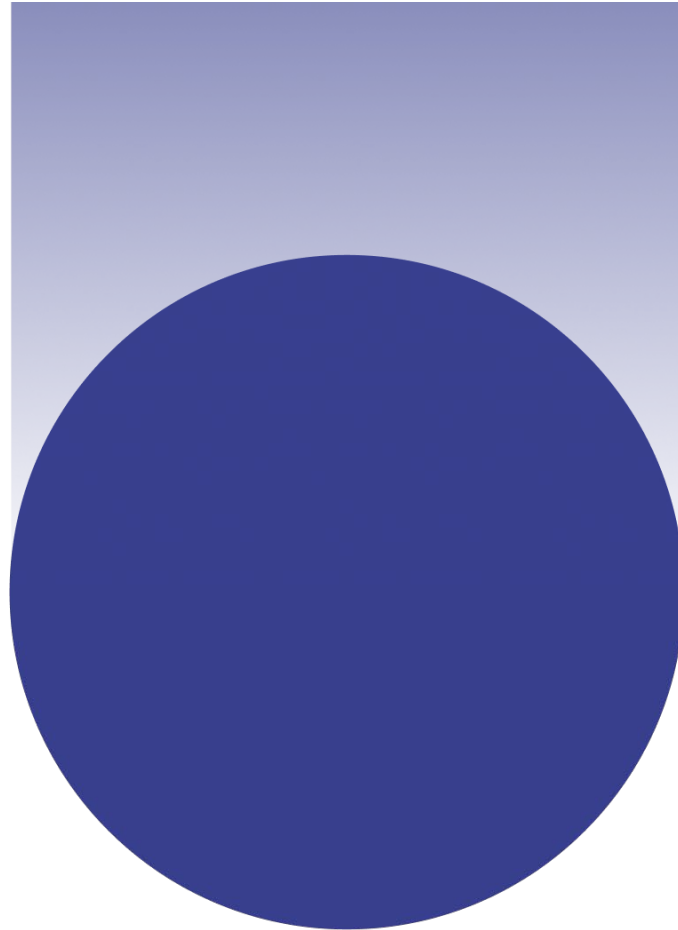
Appendices C to O can be found on the EEF website.

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at <https://educationendowmentfoundation.org.uk>



Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

<https://educationendowmentfoundation.org.uk>

 [@EducEndowFoundn](https://twitter.com/EducEndowFoundn)

 [Facebook.com/EducEndowFoundn](https://www.facebook.com/EducEndowFoundn)